CRAY
THE SUPERCOMPUTER COMPANY

Future Peta-Exa Scale Architectures

John Levesque
February, 2011

Outline

- Future architectural directions
- Cray XT6 and Cray XE6 compute blades
- Seastar and Gemini interconnects

Today’s Fastest System

Systems	2010
System peak	2 Pflop/s
Power	6 MW
System memory	0.3 PB
Node performance	125 GF
Node memory BW	25 GB/s
Node concurrency	12
Total Node Interconnect BW	3.5 GB/s
System size (nodes)	18,700
Total concurrency	225,000
Storage	15 PB
IO	0.2 TB
MTTI	days

Today’s Fastest System

Systems	2010
System peak	2 Pflop/s
Power	6 MW
System memory	0.3 PB
Node performance	125 GF
Node memory BW	25 GB/s
Node concurrency	12
Total Node Interconnect BW	3.5 GB/s
System size (nodes)	18,700
Total concurrency	225,000
Storage	15 PB
IO	0.2 TB
MTTI	days

Moving forward to Exascale…

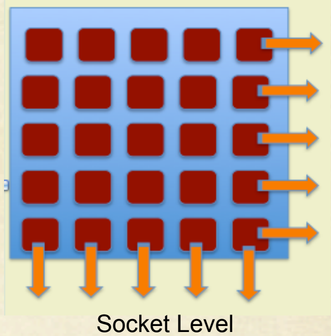
The claim is that an Exascale system can cost at most $200M and consume no more than 20 MW.

Potential System Architecture with a cap of $200M and 20MW

Systems	2010	2018	Difference Today & 2018
System peak	2 Pflop/s	1 Eflop/s	O(1000)
Power	6 MW	~20 MW	
System memory	0.3 PB	32 - 64 PB [.03 Bytes/Flop]	O(100)
Node performance	125 GF	1,2 or 15TF	O(10) – O(100)
Node memory BW	25 GB/s [.20 Bytes/Flop]	2 - 4TB/s [.002 Bytes/Flop]	O(100)
Node concurrency	12	O(1k) or 10k	O(100) – O(1000)
Total Node Interconnect BW	3.5 GB/s	200-400GB/s (1:4 or 1:8 from memory BW)	O(100)
System size (nodes)	18,700	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	225,000	O(billion) [O(10) to O(100) for latency hiding]	O(10,000)
Storage	15 PB	500-1000 PB (>10x system memory is min)	O(10) – O(100)
IO	0.2 TB	60 TB/s (how long to drain the machine)	O(100)
MTTI	days	O(1 day)	- O(10)

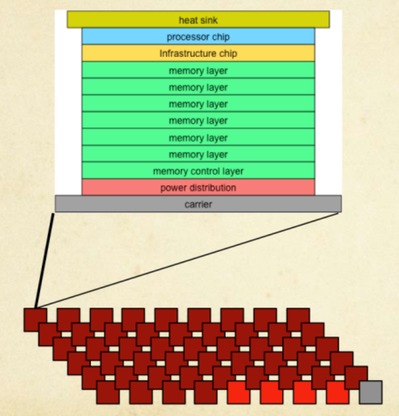
Exascale (10^{18} Flop/s) Systems: Two possible paths

- Light weight processors (think BG/P)
  - ~1 GHz processor (10^9)
  - ~1 Kilo cores/socket (10^3)
  - ~1 Mega sockets/system (10^6)



Socket Level
Cores scale-out for planar geometry

- Hybrid system (think GPU based)
  - ~1 GHz processor (10^9)
  - ~10 Kilo FPUs/socket (10^4)
  - ~100 Kilo sockets/system (10^5)



Node Level
3D packaging

Short Term Petascale Systems – Node Architecture

	Cores on the node	Total threading	Vector Length	Programming Model
Blue Waters	16	32	8	OpenMP/MPI/ Vector
Blue Gene Q	16	32	8	OpenMP/MPI/ Vector
Magna-Cours	24	24	4	OpenMP/MPI/ Vector
OLCF3-Acc	32	32 (768*)	16	Threads/ Cuda/Vector
Intel MIC	32	128	8	OpenMP/MPI/ Vector
Jaguar & Kraken	12	12	2	OpenMP/MPI/ Vector

* Nvidia allows oversubscription to SIMT units

Hybrid Multi-core Architecture

- Massively Parallel System with high powered nodes that exhibit
  - Multiple levels of parallelism
    - Shared Memory parallelism on the node
    - SIMD vector units on each core or thread
  - Potentially disparate processing units
    - Host with conventional X86 architecture
    - Accelerator with highly parallel – SIMD units
  - Potentially disparate memories
    - Host with conventional DDR memory
    - Accelerator with high bandwidth memory

Hybrid Multi-core Architecture

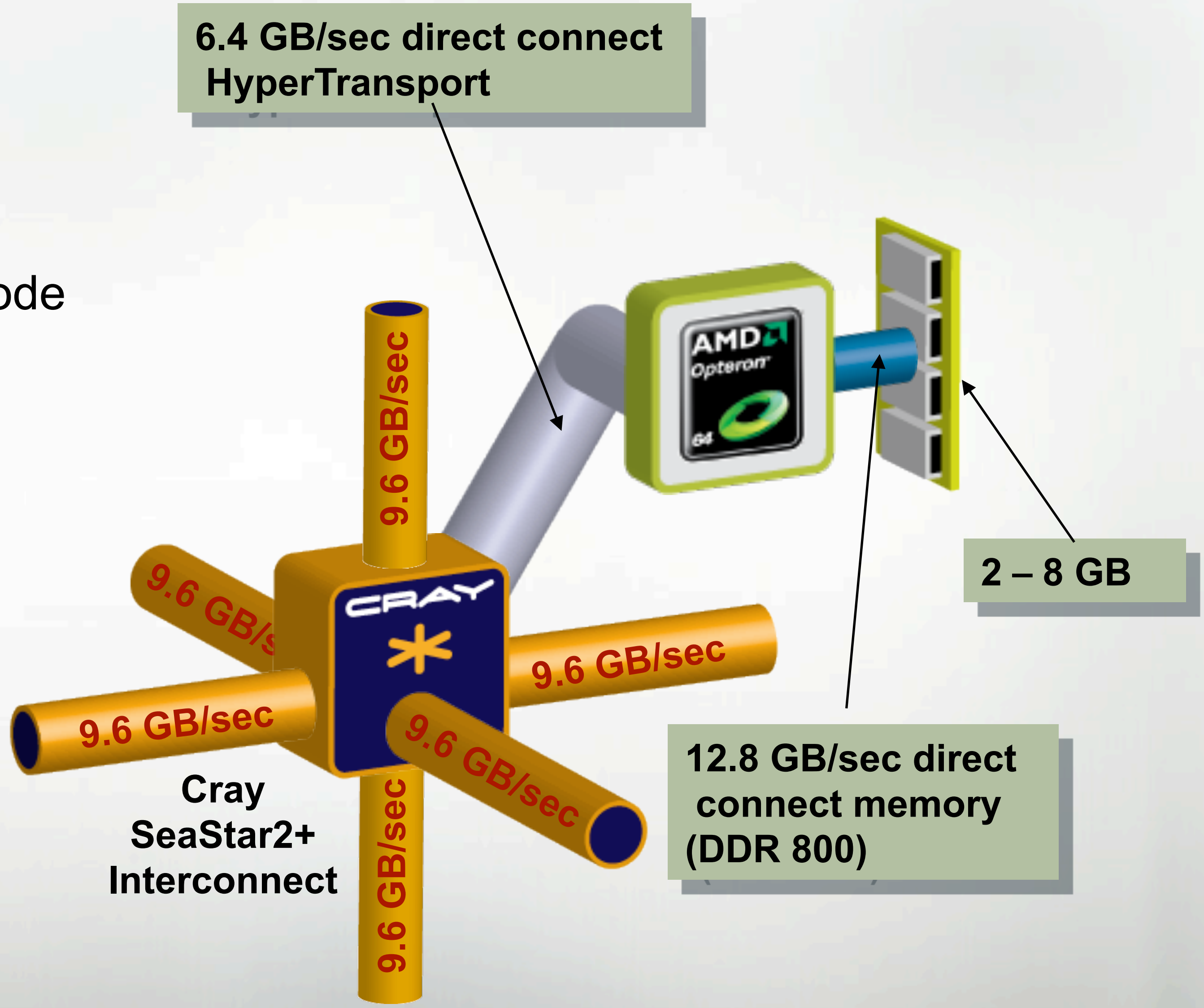
- All MPI may not be best approach
  - Memory per core will decease
  - Injection bandwidth/core will decease
  - Memory bandwidth/core will decrease
- Hybrid MPI + threading on node may be able to
  - Save Memory
  - Reduce amount of off node communication required
  - Reduce amount of memory bandwidth required

An overview of Cray XT systems

	XT3	XT4	XT5	XT6
Number of cores/socket	2	4	4-6	12
Number of cores/node	2	4	8-12	24
Clock Cycle (CC)	2.6	2.3	2.6	2.1
Number of 64 bit Results/CC	2	4	4	4
GFLOPS/Node	10.4	36.8	83.6-124.8	~200
Interconnect	Seastar 1	Seastar 2+	Seastar 2+	Gemini
Link Bandwidth GB/sec	6x2.4	6x4.8	6x4.8	10x4.7
MPI Latency microseconds	6	6	6	1.5
Messages/sec	400K	400K	400K	10M
Global Addressing	No	No	No	Yes

Cray XT4 Node

- 4-way SMP
- >35 Gflops per node
- Up to 8 GB per node
- OpenMP Support within socket



6.4 GB/sec direct connect HyperTransport

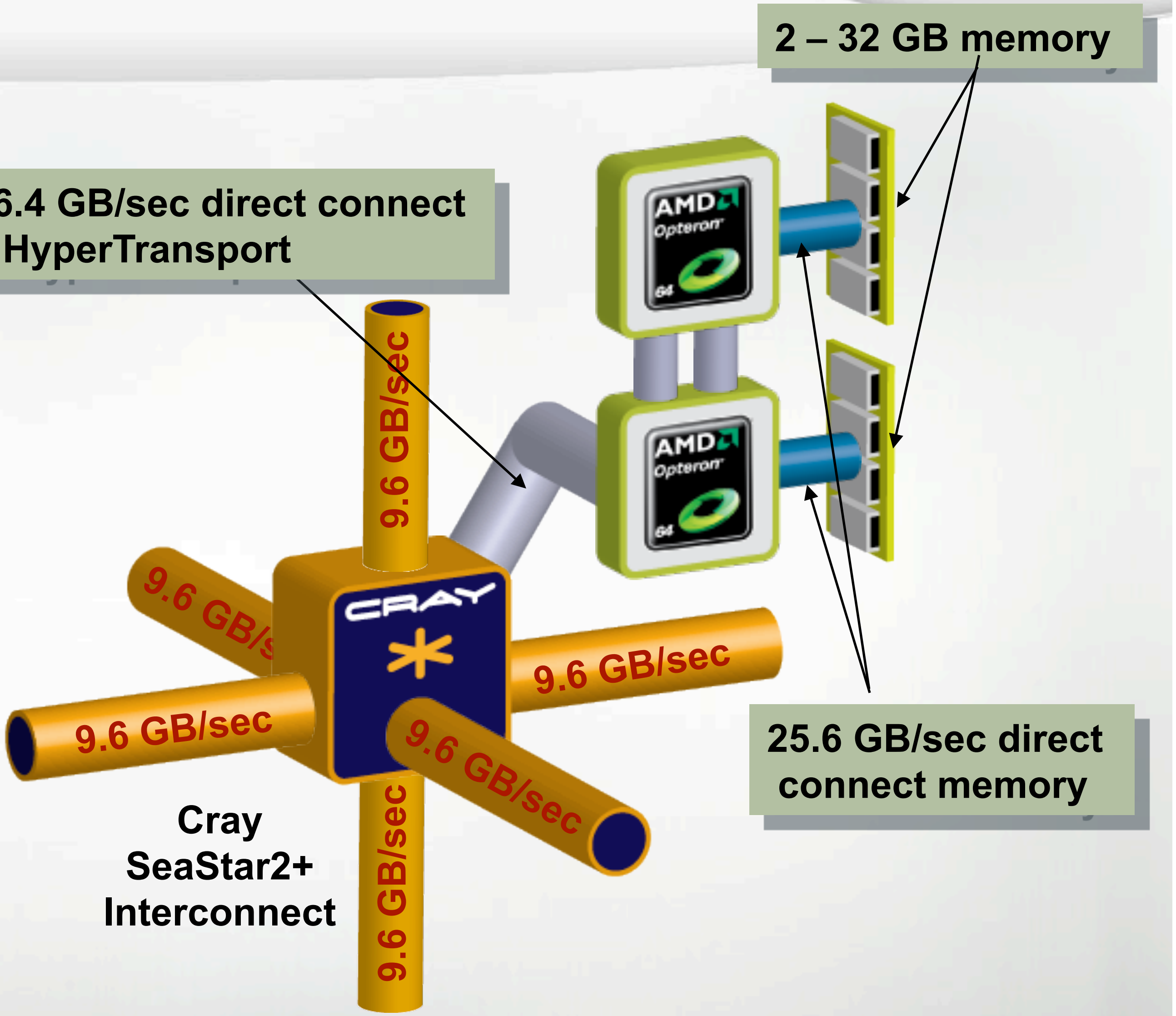
2 – 8 GB

12.8 GB/sec direct connect memory (DDR 800)

Cray SeaStar2+ Interconnect

Cray XT5 Node

- 8-way SMP
- >70 Gflops per node
- Up to 32 GB of shared memory per node
- OpenMP Support



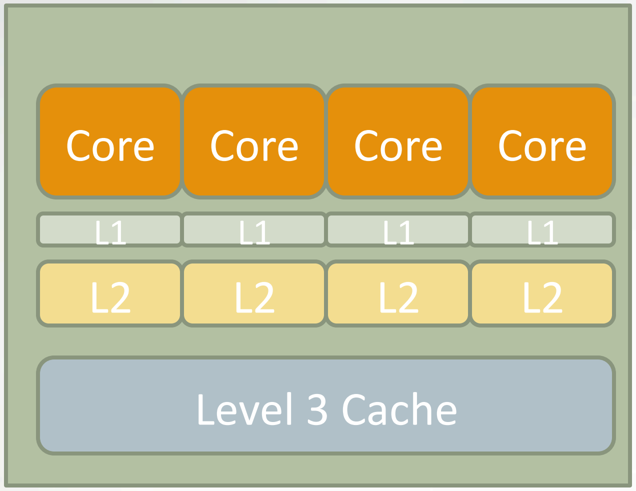
2 – 32 GB memory

6.4 GB/sec direct connect HyperTransport

Cray SeaStar2+ Interconnect

25.6 GB/sec direct connect memory

XT Hardware Socket



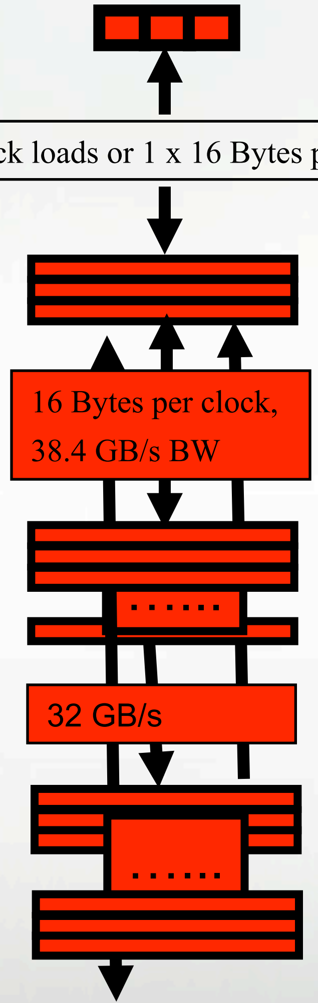
Budapest	Barcelona	Shanghai	Istanbul
4 FPS/CC	4 FPS/CC	4 FPS/CC	4 FPS/CC
64KB	64KB	64KB	64KB
512KB	512KB	512KB	512KB
2048KB	2048KB	6144KB	6144KB

Simplified memory hierachy on the Quad Core AMD Opteron – Quad Core

registers

16 SSE2 128-bit registers
16 64 bit registers

2 x 16 Bytes per clock loads or 1 x 16 Bytes per clock store, (76.8 GB/s or 38.4 GB/s on 2.4 Ghz)



L1 data cache

L2 cache

Shared L3 cache

- 64 Byte cache line
- complete data cache lines are loaded from main memory, if not in L2 or L3 cache
- if L1 data cache needs to be refilled, then storing back to L2 cache, if L2 needs to be refilled, storing back to L3
- Items in L1 and L2 are exclusive, L3 is "sharing aware"
- write back cache: data offloaded from L1 data cache are stored here first until they are flushed out to main memory
- Pre-fetching can go to L1 or L2 cache

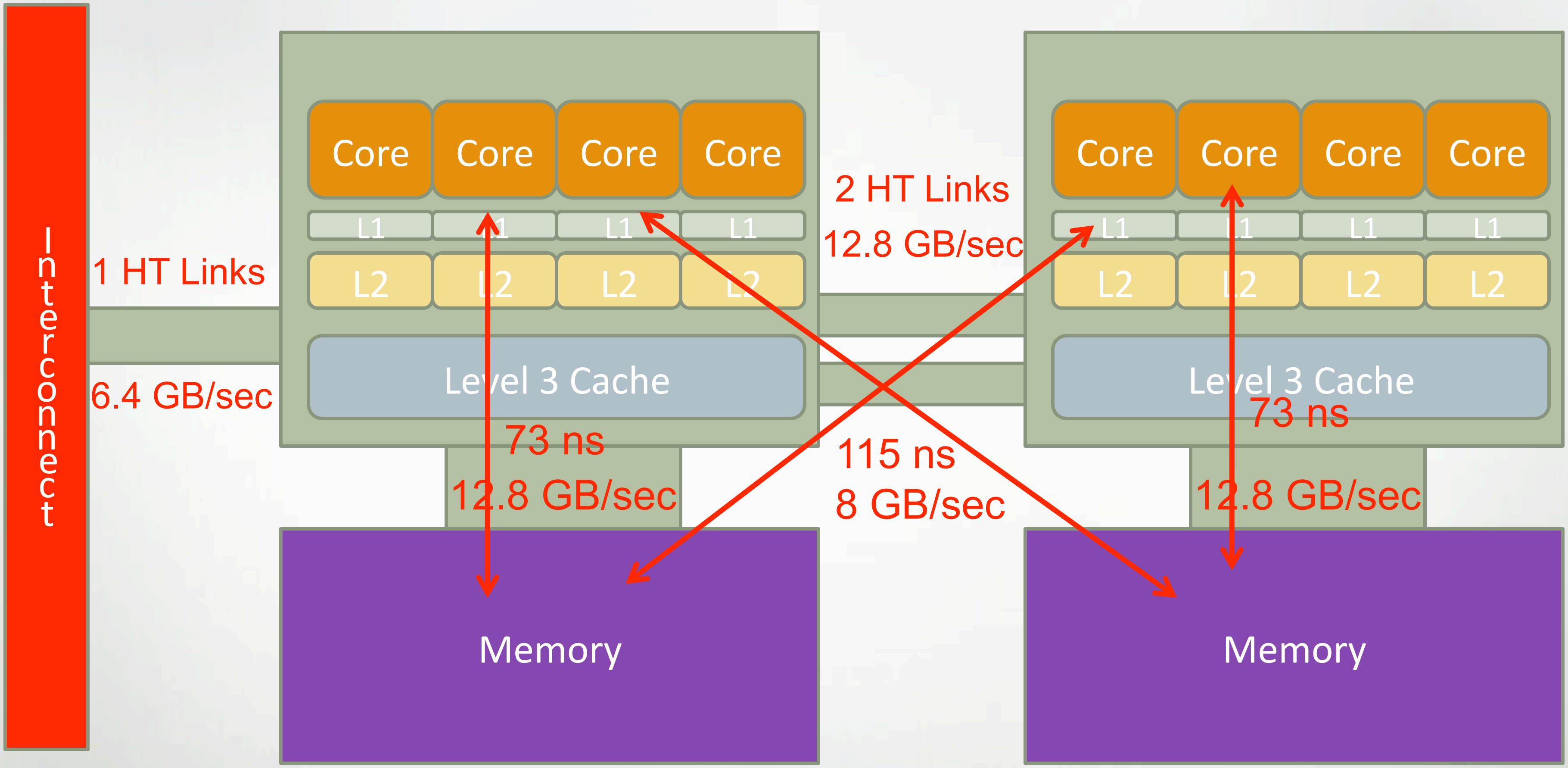
16 Bytes wide => 12.8 GB/s for DDR2-800, 73ns

Main memory

8GB/s over coherent Hyper Transport, 115ns

Remote memory

XT Hardware Node



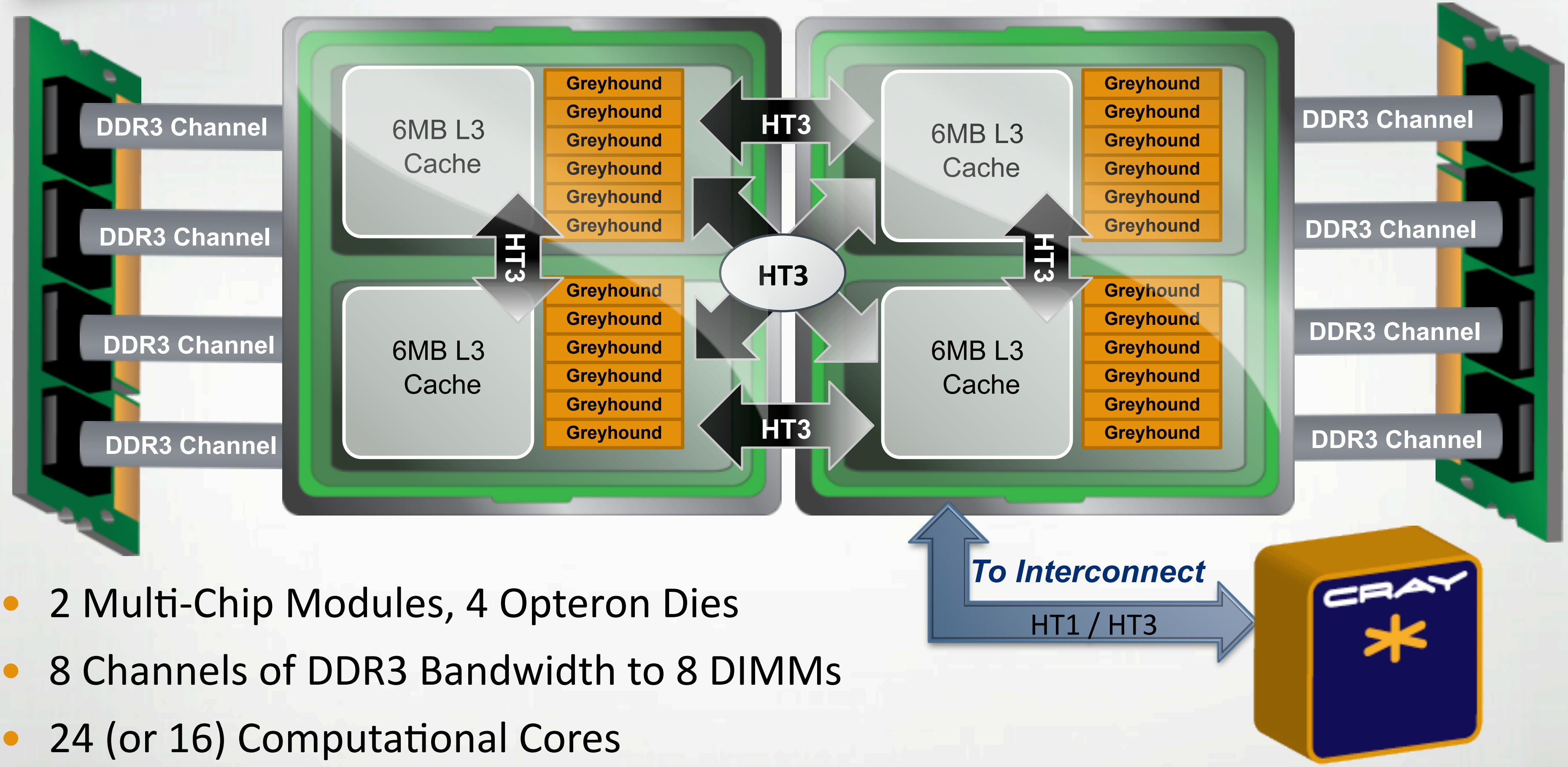
What are the strengths and bottlenecks of the hardware

- Strengths
  - Upgradability
  - Scalability of interconnects
  - Increased node performance – need to use fewer nodes to achieve same performance
  - Global addressing with Gemini
    - Adds ability to use PGAS – UPC And CAF to program around some of the weaknesses

What are the strengths and bottlenecks of the hardware

- Bottlenecks
  - Memory bandwidth will never be enough to support all the cores
    - Need to think about programming around this
      - PGAS – UPC and CAF
      - OpenMP
  - Injection Bandwidth will never be enough to support all the cores
    - Need to think about programming around this
      - PGAS – UPC and CAF
      - OpenMP
  - Global Bandwidth will never be enough to support ALL to ALLs across all of the cores
    - Need to think about programming around this
      - PGAS – UPC and CAF
      - OpenMP

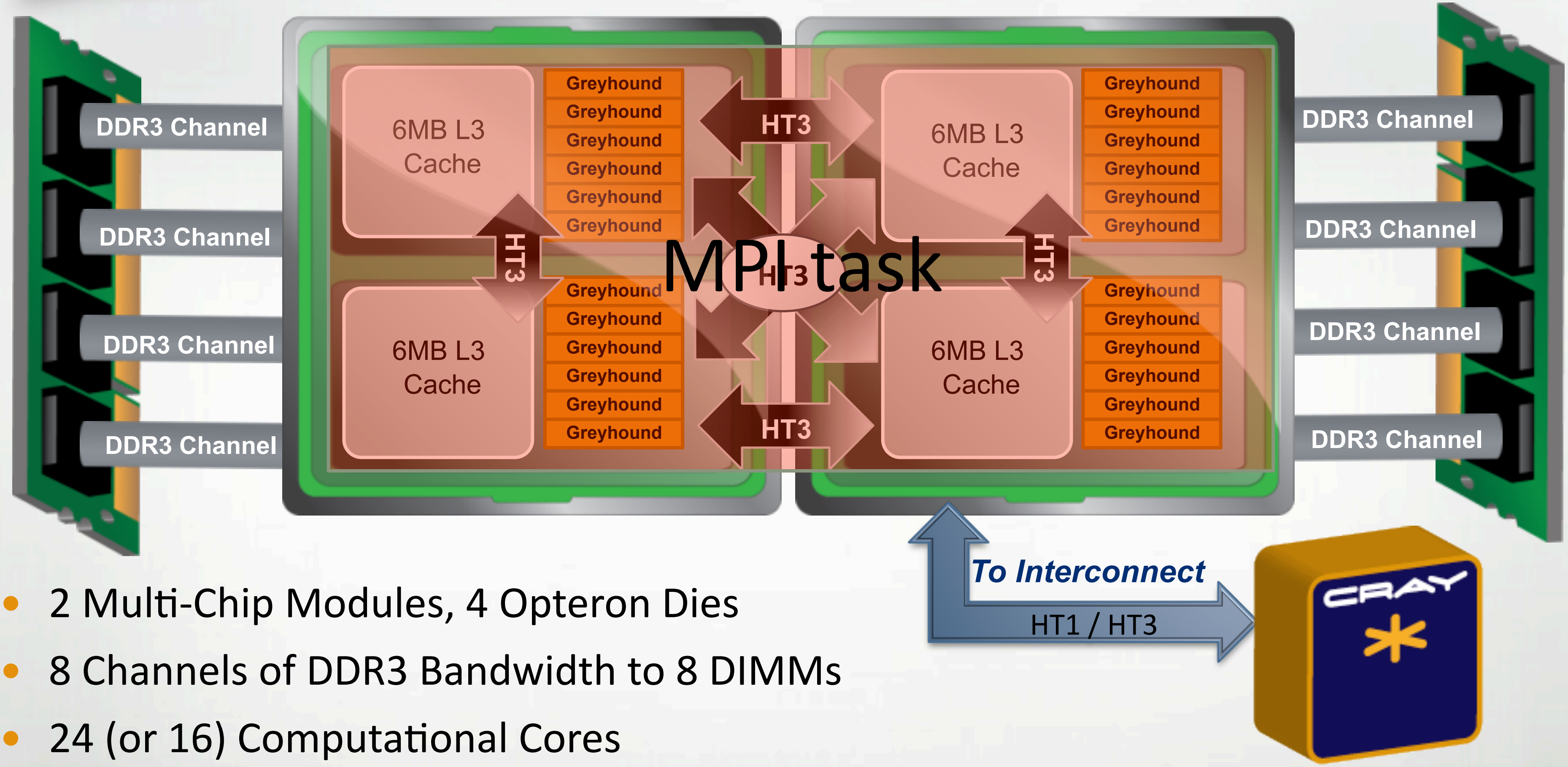
XE6 Node Details:
24-core Magny Cours



- 2 Multi-Chip Modules, 4 Opteron Dies
- 8 Channels of DDR3 Bandwidth to 8 DIMMs
- 24 (or 16) Computational Cores
  - 64 KB L1 and 512 KB L2 caches for each core
  - 6 MB of shared L3 cache on each die
- Dies are fully connected with HT3
- Snoop Filter Feature Allows 4 Die SMP to scale well

XE6 Node Details: 24-core Magny Cours

Run using 1 MPI task on the node

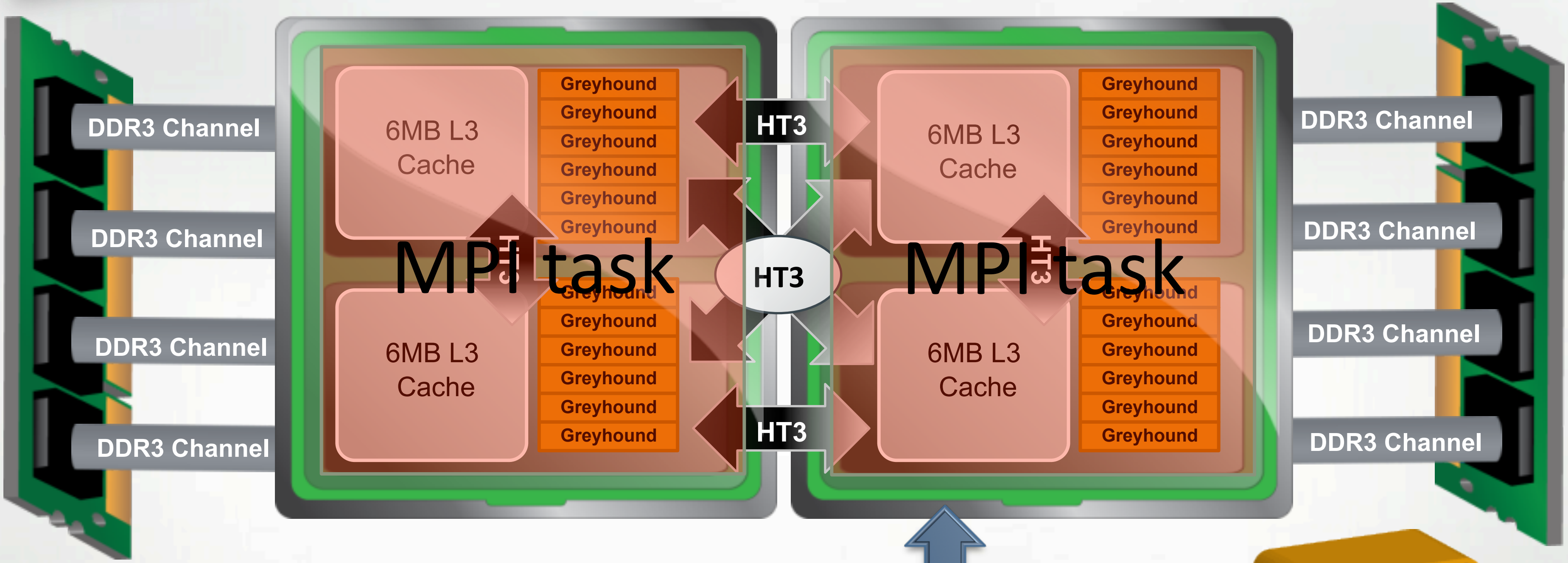


- 2 Multi-Chip Modules, 4 Opteron Dies
- 8 Channels of DDR3 Bandwidth to 8 DIMMs
- 24 (or 16) Computational Cores
  - 64 KB L1 and 512 KB L2 caches for each core
  - 6 MB of shared L3 cache on each die
- Dies are fully connected with HT3
- Snoop Filter Feature Allows 4 Die SMP to scale well

Use OpenMP across all 24 cores

XE6 Node Details: 24-core Magny Cours

Run using 2 MPI tasks on the node
One on Each Die



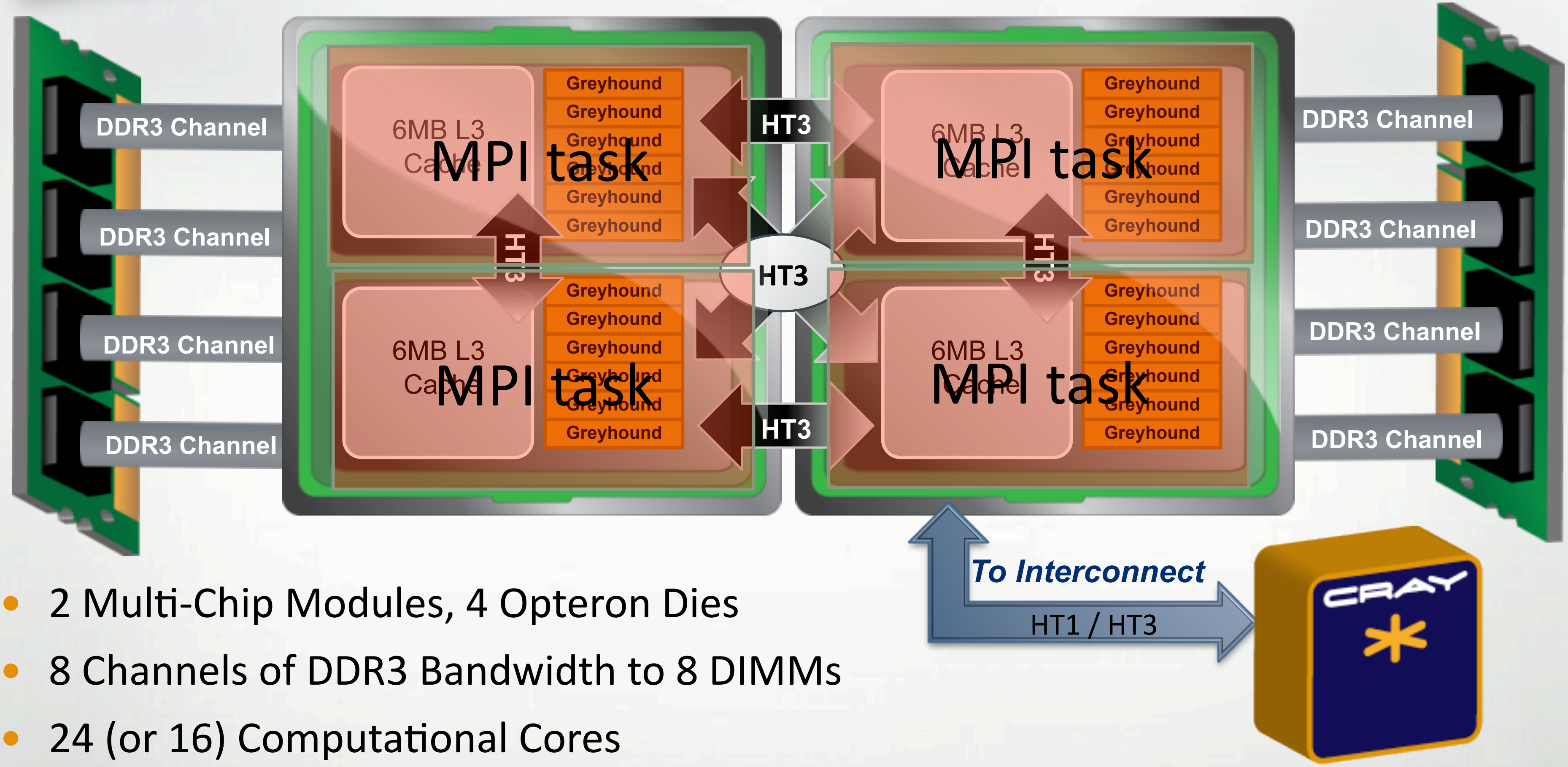
To Interconnect
HT1 / HT3
CRAY

- 2 Multi-Chip Modules, 4 Opteron Dies
- 8 Channels of DDR3 Bandwidth to 8 DIMMs
- 24 (or 16) Computational Cores
  - 64 KB L1 and 512 KB L2 caches for each core
  - 6 MB of shared L3 cache on each die
- Dies are fully connected with HT3
- Snoop Filter Feature Allows 4 Die SMP to scale well

Use OpenMP across all 12 cores in the Die

XE6 Node Details:
24-core Magny Cours

Run using 4 MPI tasks on the node
One on Each Socket



- 2 Multi-Chip Modules, 4 Opteron Dies
- 8 Channels of DDR3 Bandwidth to 8 DIMMs
- 24 (or 16) Computational Cores
  - 64 KB L1 and 512 KB L2 caches for each core
  - 6 MB of shared L3 cache on each die
- Dies are fully connected with HT3
- Snoop Filter Feature Allows 4 Die SMP to scale well

Use OpenMP across all 6 cores in the Socket

Proposed Programming Paradigm for Hybrid Multi-core

- MPI or PGAS between nodes and/or sockets
- OpenMP, Pthreads or some other shared memory parallelism across a portion of the cores on the node
- Vectorization to utilize the SSE# or SIMD units on the cores

Why should you be concerned

- Node structure will be the same for all system sizes
  - Exascale
  - Petascale
  - Terascale

Will all use the same node

Check out Website for Future Developments

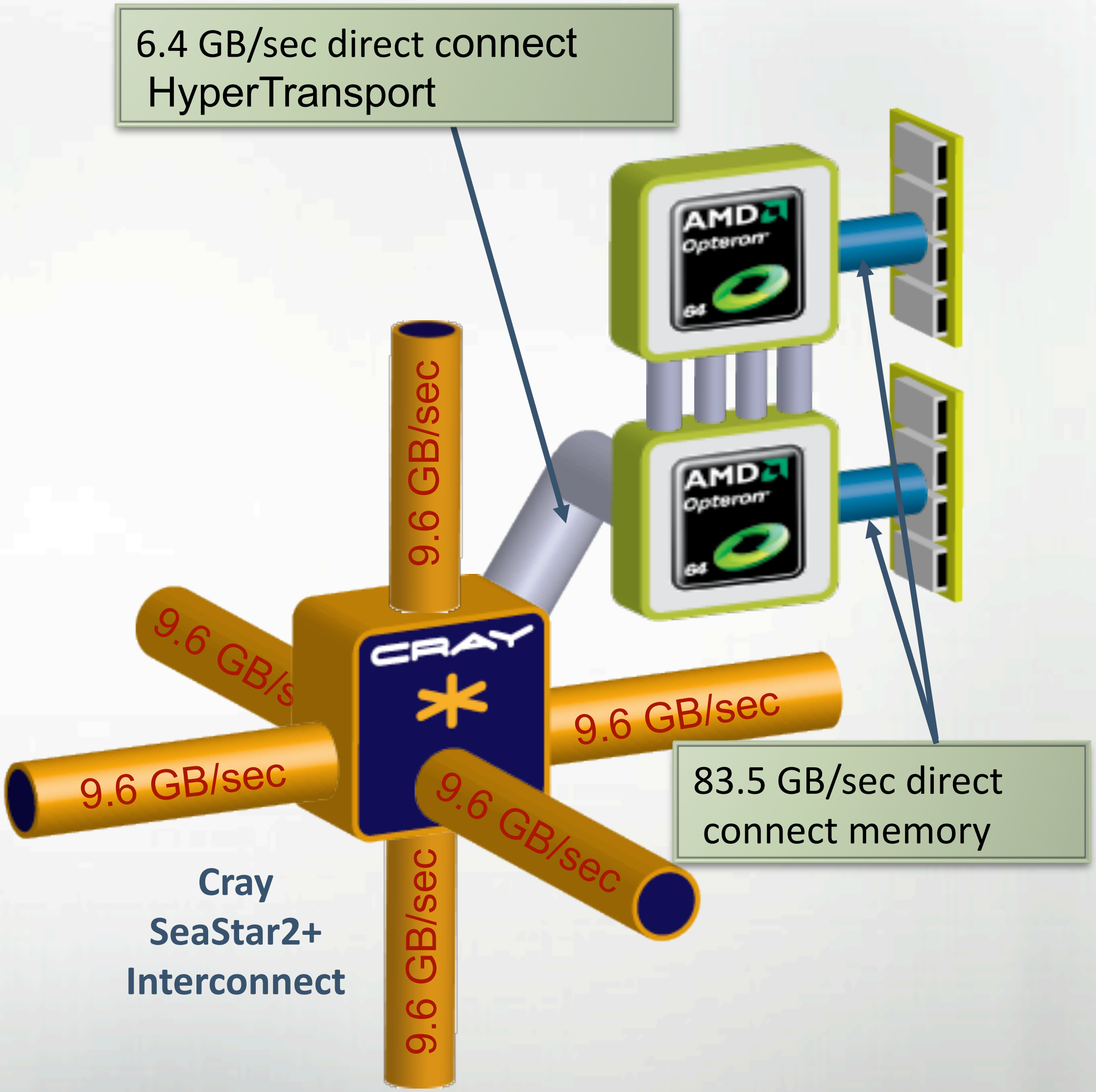
http://hybridmulticore.com/

CRAY
THE SUPERCOMPUTER COMPANY

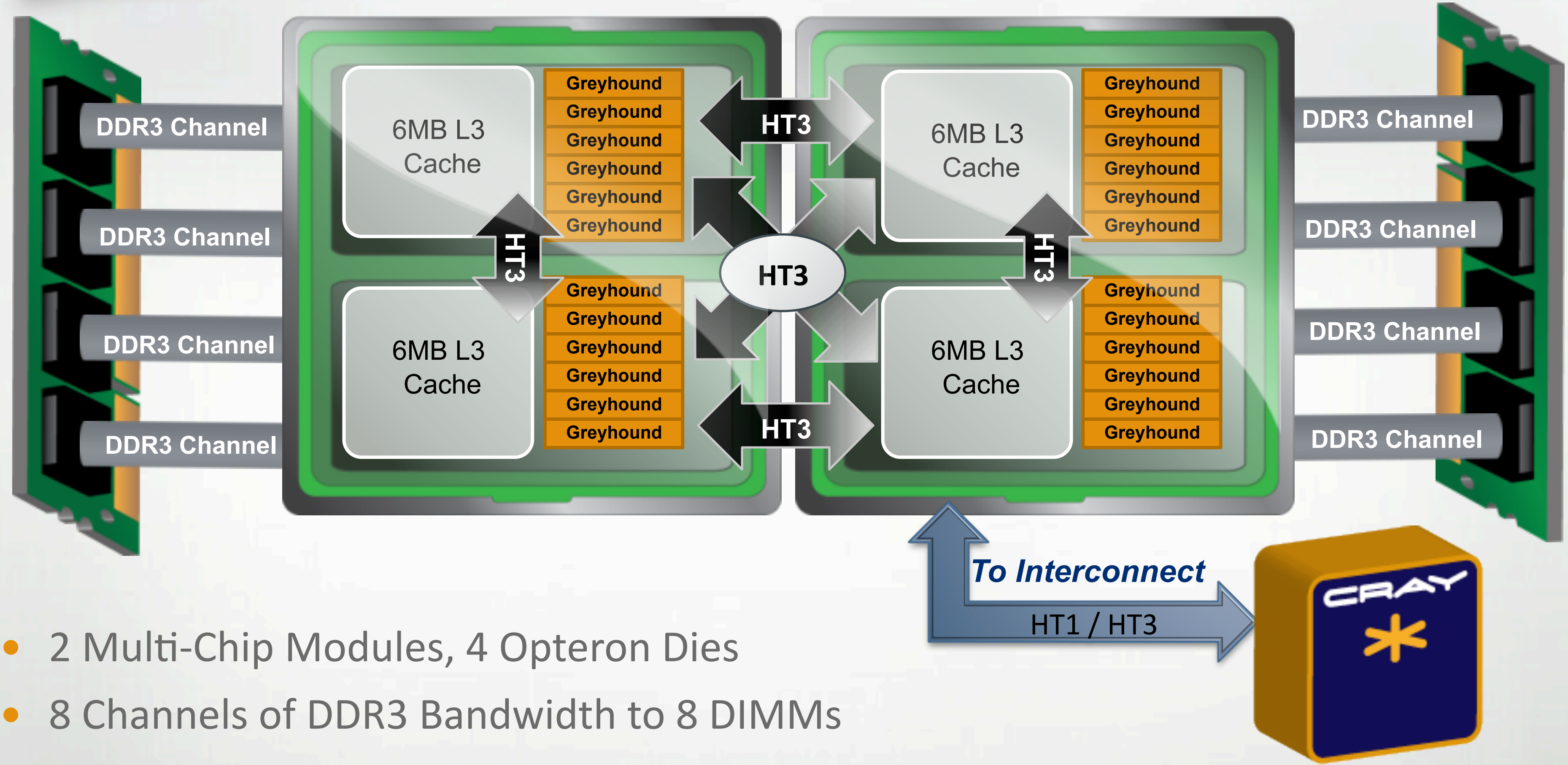
Cray XT6

Cray XT6 (Or XT6m) Node

Characteristics	
Number of Cores	16 or 24 (MC)
Peak Performance MC-8 (2.4)	153 Gflops/sec
Peak Performance MC-12 (2.2)	211 Gflops/sec
Memory Size	32 or 64 GB per node
Memory Bandwidth	83.5 GB/sec

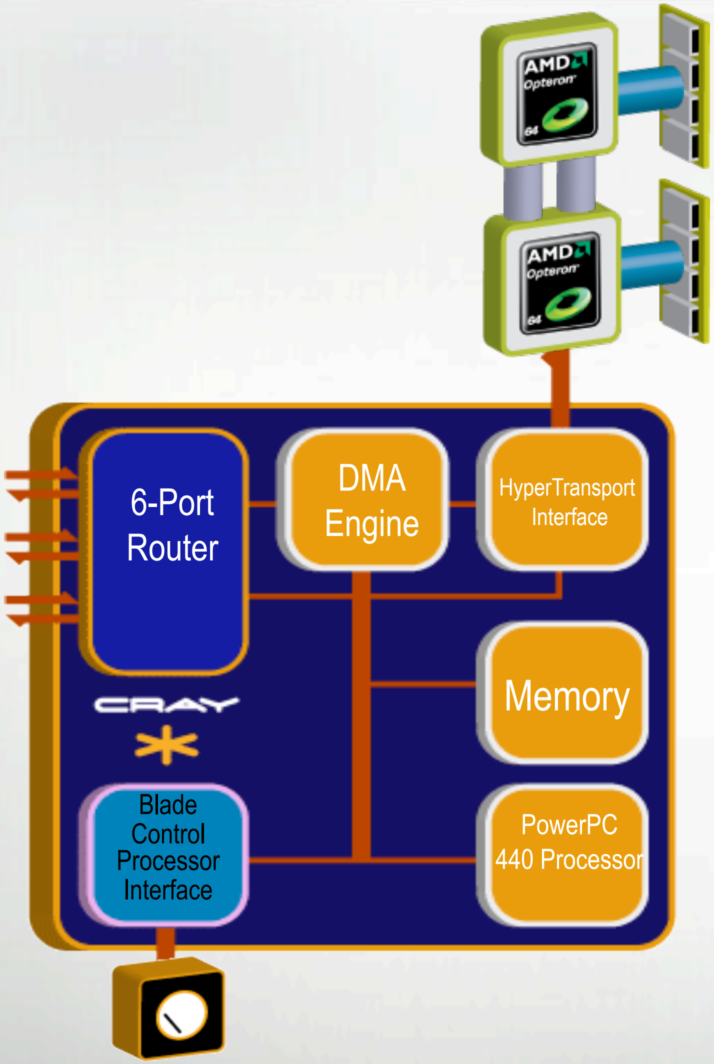

6.4 GB/sec direct connect HyperTransport
AMD Opteron 64
AMD Opteron 64
9.6 GB/sec
9.6 GB/sec
9.6 GB/sec
9.6 GB/sec
9.6 GB/sec
9.6 GB/sec
CRAY
83.5 GB/sec direct connect memory
Cray SeaStar2+ Interconnect

XT6 or XE6 Node Details:
24-core Magny Cours



- 2 Multi-Chip Modules, 4 Opteron Dies
- 8 Channels of DDR3 Bandwidth to 8 DIMMs
- 24 (or 16) Computational Cores, 24 MB of L3 cache
- Dies are fully connected with HT3
- Snoop Filter Feature Allows 4 Die SMP to scale well

Cray SeaStar2+ Interconnect

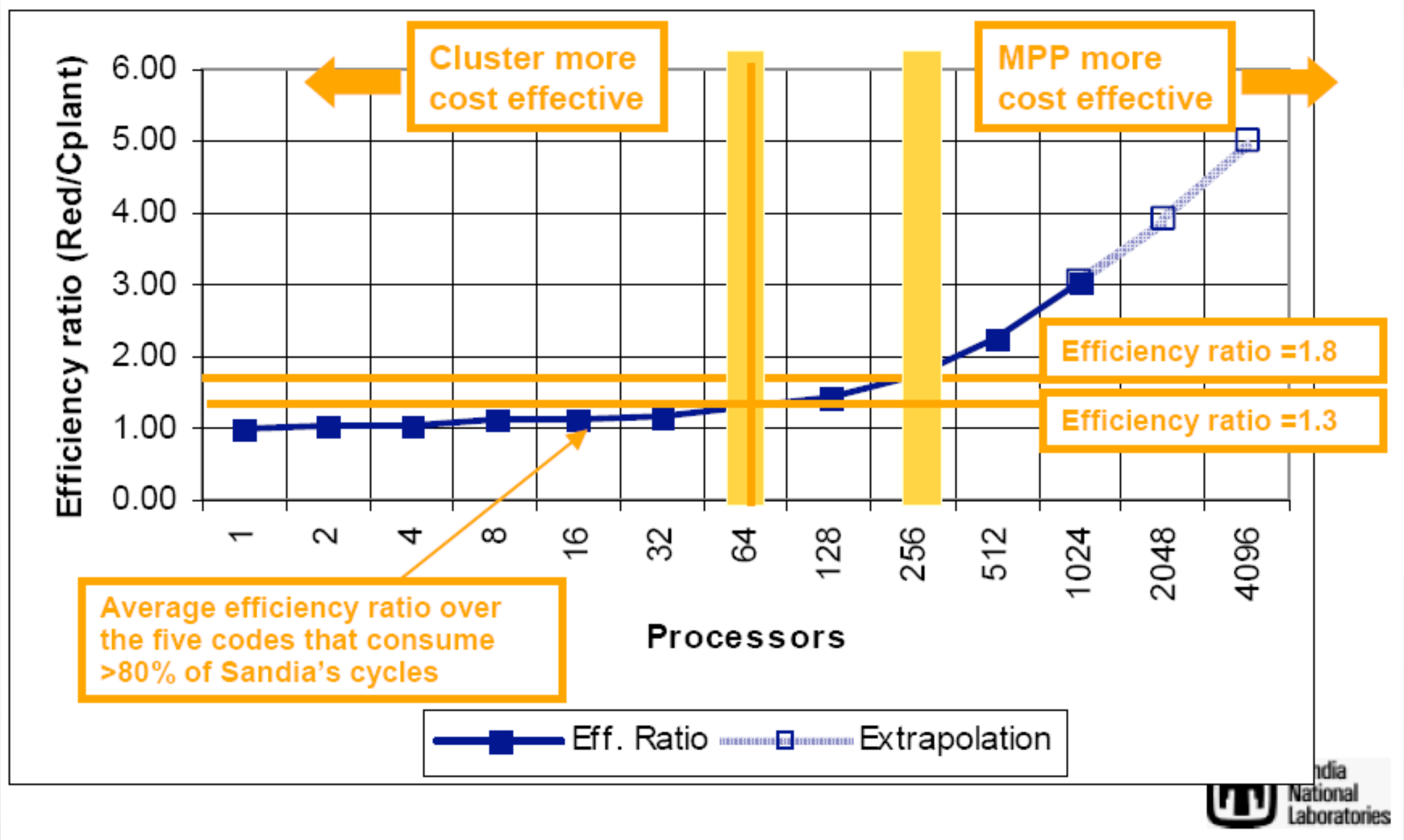


Now Scaled to 225,000 cores

- **Cray XT6 systems ship with the SeaStar2+ interconnect**
- Custom ASIC
- **Integrated NIC / Router**
- MPI offload engine
- Connectionless Protocol
- Link Level Reliability
- Proven scalability to 225,000 cores

Why Custom Interconnects ?

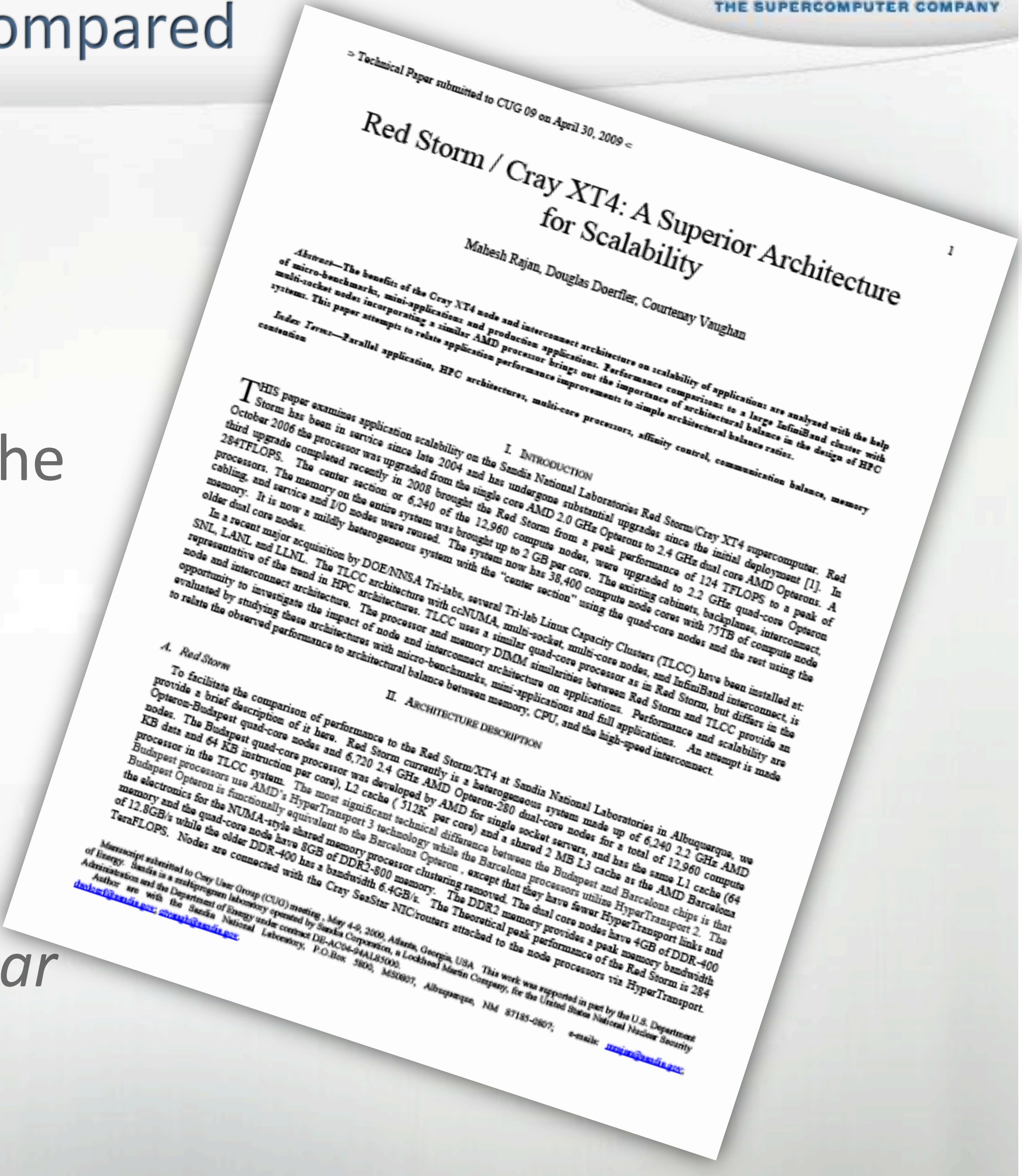
- We have been keeping a close eye on progress with infiniband but we still believe custom interconnects are necessary for HPC



- What follows is a customer performance example showing the difference
- And a usage pattern we see on NSF systems…

SeaStar and Infiniband Compared

- SeaStar-based XT5 systems are the first systems to scale application codes to the Petaflop level
- SeaStar provides scalability advantages at much smaller processor counts, *particularly on irregular communication patterns*



> Technical Paper submitted to CUG 09 on April 30, 2009 <

1

Red Storm / Cray XT4: A Superior Architecture for Scalability

Mahesh Rajan, Douglas Doerfler, Courtenay Vaughan

***Abstract*—The benefits of the Cray XT4 node and interconnect architecture on scalability of applications are analyzed with the help of micro-benchmarks, mini-applications and production applications. Performance comparisons to a large InfiniBand cluster with multi-socket nodes incorporating a similar AMD processor brings out the importance of architectural balance in the design of HPC systems. This paper attempts to relate application performance improvements to simple architectural balance ratios.**

***Index Terms*—Parallel application, HPC architectures, multi-core processors, affinity control, communication balance, memory contention**

I. Introduction

THIS paper examines application scalability on the Sandia National Laboratories Red Storm/Cray XT4 supercomputer. Red Storm has been in service since late 2004 and has undergone substantial upgrades since the initial deployment [1]. In October 2006 the processor was upgraded from the single core AMD 2.0 GHz Opterons to 2.4 GHz dual core AMD Opterons. A third upgrade completed recently in 2008 brought the Red Storm from a peak performance of 124 TFLOPS to a peak of 284TFLOPS. The center section or 6,240 of the 12,960 compute nodes, were upgraded to 2.2 GHz quad-core Opteron processors. The memory on the entire system was brought up to 2 GB per core. The existing cabinets, backplanes, interconnect, cabling, and service and I/O nodes were reused. The system now has 38,400 compute node cores with 75TB of compute node memory. It is now a mildly heterogeneous system with the "center section" using the quad-core nodes and the rest using the older dual core nodes.

In a recent major acquisition by DOE/NNSA Tri-labs, several Tri-lab Linux Capacity Clusters (TLCC) have been installed at: SNL, LANL and LLNL. The TLCC architecture with ccNUMA, multi-socket, multi-core nodes, and InfiniBand interconnect, is representative of the trend in HPC architectures. TLCC uses a similar quad-core processor as in Red Storm, but differs in the node and interconnect architecture. The processor and memory DIMM similarities between Red Storm and TLCC provide an opportunity to investigate the impact of node and interconnect architecture on applications. Performance and scalability are evaluated by studying these architectures with micro-benchmarks, mini-applications and full applications. An attempt is made to relate the observed performance to architectural balance between memory, CPU, and the high-speed interconnect.

II. Architecture Description

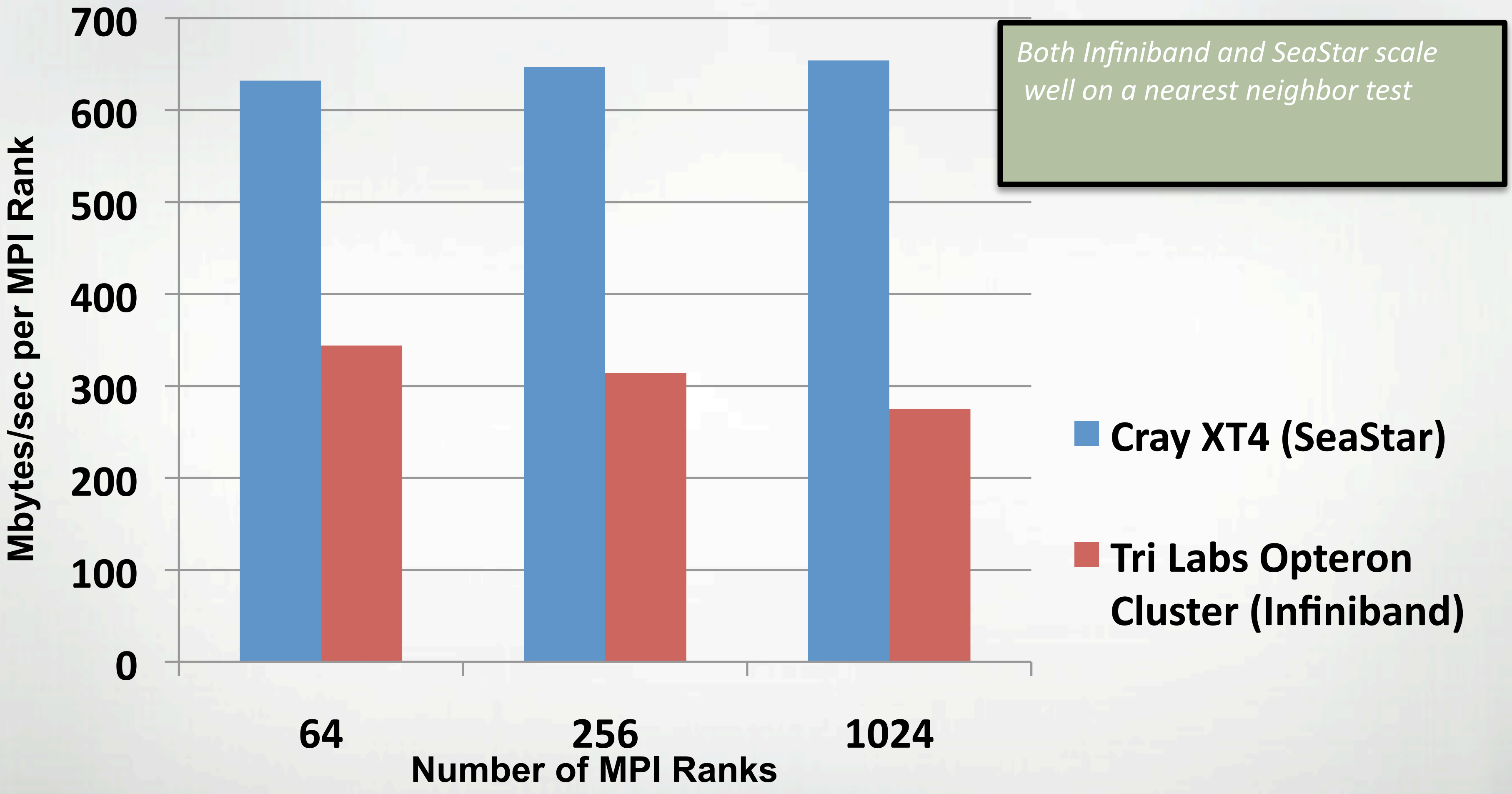
A. Red Storm

To facilitate the comparison of performance to the Red Storm/XT4 at Sandia National Laboratories in Albuquerque, we provide a brief description of it here. Red Storm currently is a heterogeneous system made up of 6,240 2.2 GHz AMD Opteron-Budapest quad-core nodes and 6,720 2.4 GHz AMD Opteron-280 dual-core nodes for a total of 12,960 compute nodes. The Budapest quad-core processor was developed by AMD for single socket servers, and has the same L1 cache (64 KB data and 64 KB instruction per core), L2 cache (512K per core) and a shared 2 MB L3 cache as the AMD Barcelona processor in the TLCC system. The most significant technical difference between the Budapest and Barcelona chips is that Budapest processors use AMD's HyperTransport 3 technology while the Barcelona processors utilize HyperTransport 2. The Budapest Opteron is functionally equivalent to the Barcelona Opteron , except that they have fewer HyperTransport links and the electronics for the NUMA-style shared memory processor clustering removed. The dual core nodes have 4GB of DDR-400 memory and the quad-core node have 8GB of DDR2-800 memory. The DDR2 memory provides a peak memory bandwidth of 12.8GB/s while the older DDR-400 has a bandwidth 6.4GB/s. The Theoretical peak performance of the Red Storm is 284 TeraFLOPS. Nodes are connected with the Cray SeaStar NIC/routers attached to the node processors via HyperTransport.

Manuscript submitted to Cray User Group (CUG) meeting , May 4-9, 2009, Atlanta, Georgia, USA. This work was supported in part by the U.S. Department of Energy. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States National Nuclear Security Administration and the Department of Energy under contract DE-AC04-94AL85000.

Author are with the Sandia National Laboratory, P.O.Box 5800, MS0807, Albuquerque, NM 87185-0807; e-mails: mrajan@sandia.gov; dwdoerf@sandia.gov; ctvaugh@sandia.gov.

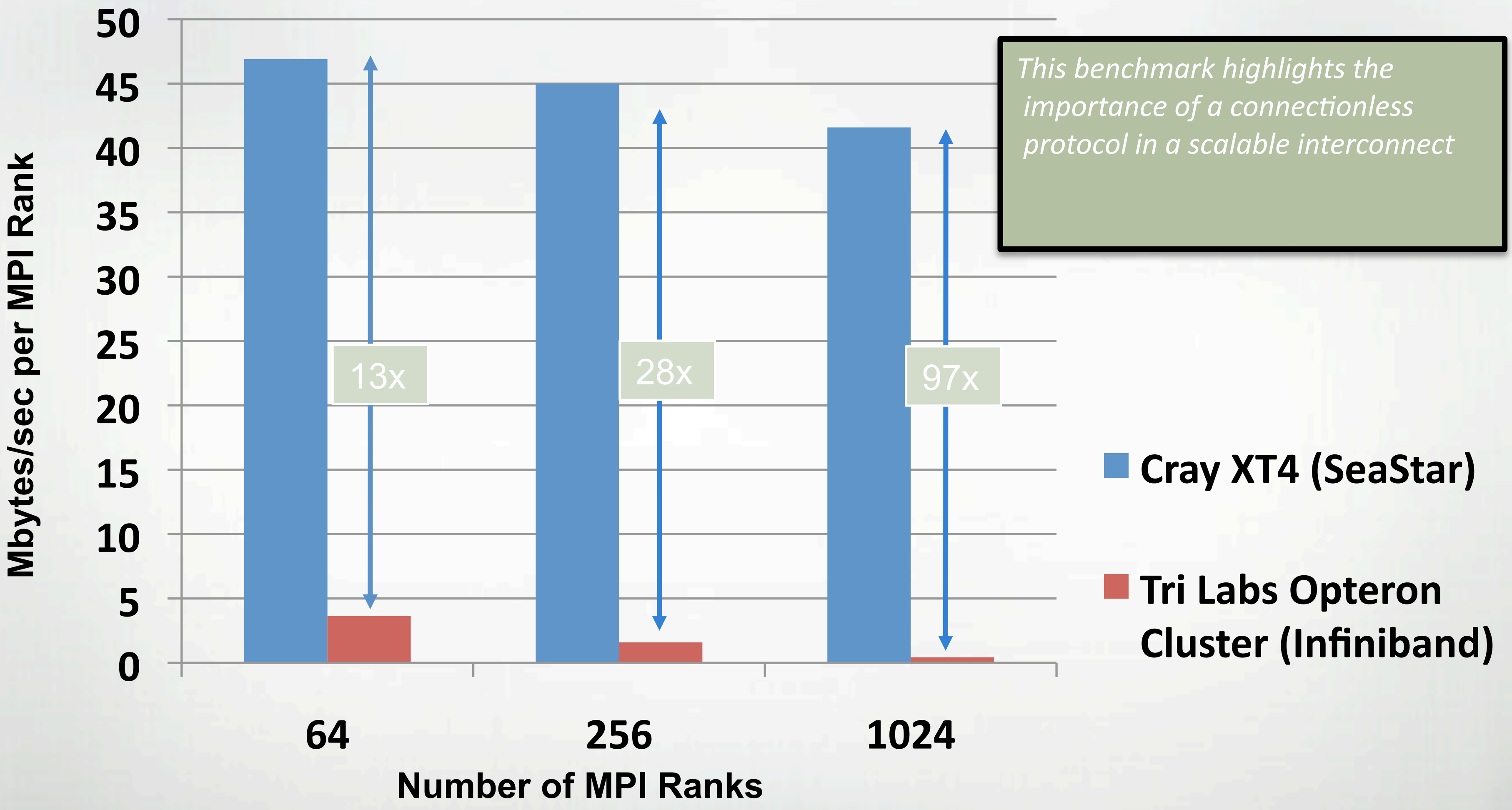
Nearest Neighbor MPI Benchmark (Bucket Brigade, Large Messages)



Both Infiniband and SeaStar scale well on a nearest neighbor test

Data from *Red Storm / Cray XT4: A Superior Architecture for Scalability* by ***Mahesh Rajan, Doug Doerfler, Courtenay Vaughan, Sandia National Laboratory*** **- Presented at Cray User Group, May 4-9, 2009**

Random Message MPI Benchmark Test (Short Messages)

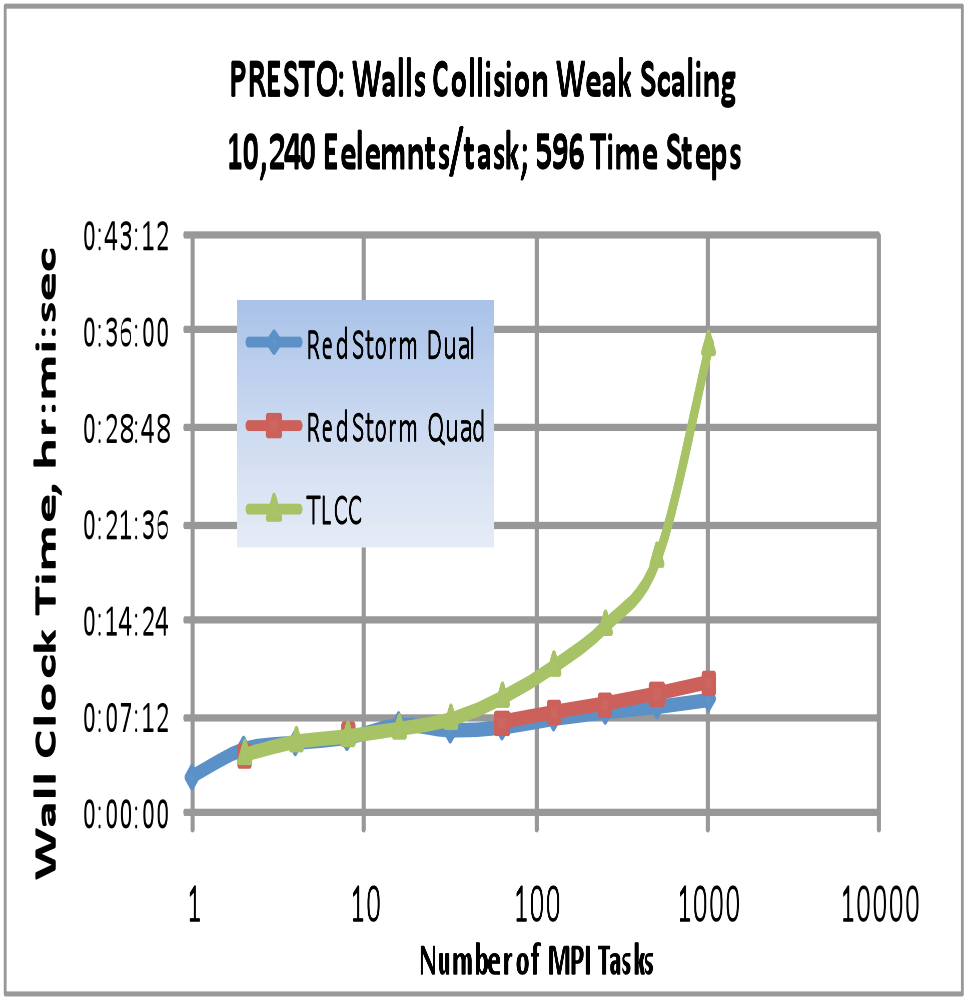


This benchmark highlights the importance of a connectionless protocol in a scalable interconnect

Data from *Red Storm / Cray XT4: A Superior Architecture for Scalability* by ***Mahesh Rajan, Doug Doerfler, Courtenay Vaughan, Sandia National Laboratory*** **- Presented at Cray User Group, May 4-9, 2009**

SIERRA/Presto – Weak Scaling

- ➢Explicit Lagrangian mechanics with contact
- ➢Model: Two sets of brick-walls colliding
- ➢Weak scaling analysis with 80 bricks /PE, each discretized with 4x4x8 elements
- ➢ Contact algorithm communications dominates the run time
- ➢ ***The rapid increase in run time after 64 processors on TLCC can be directly related to the poor performance on TLCC for random small-to-medium size messages***
- ➢ TLCC/Quad run time ratio at 1024 is 4X.

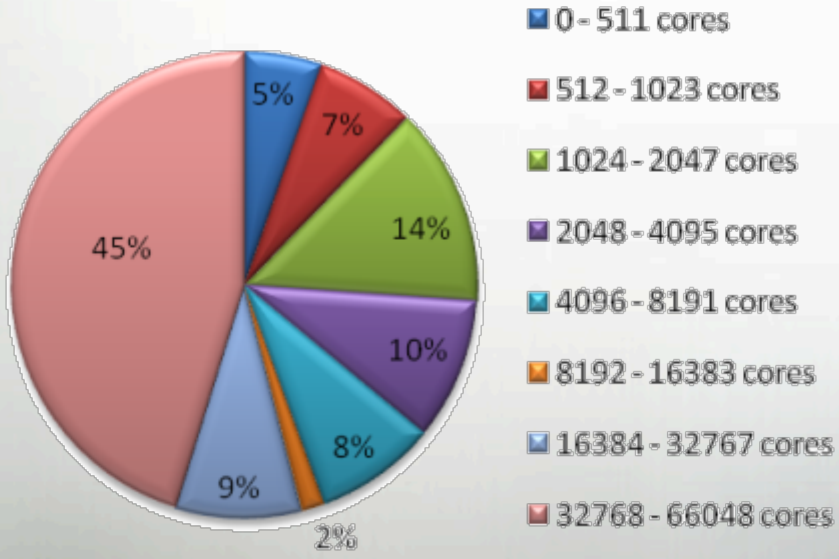


Usage Pattern – UT's Kraken Machine

Award(U. Tennessee/ORNL)	Sep, 2007
Cray XT3: 7K cores, 40 TF	Jun, 2008
Cray XT4: 18K cores,166 TF	Aug 18, 2008
Cray XT5: 65K cores, 600 TF	Feb 2, 2009
Cray XT5+: ~100K cores, 1 PF	Oct, 2009

Kraken and Krakettes!

XT5 CPU Usage by Core-Count



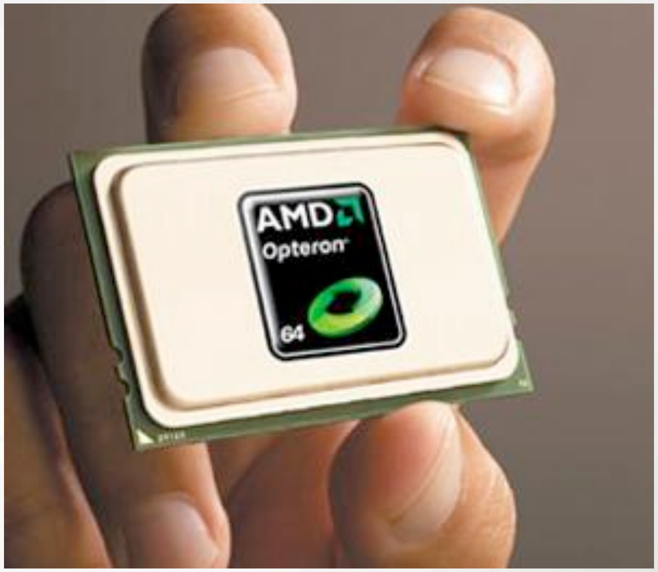
Proprietary Cray Interconnect, SeaStar2, provides excellent scaling, with numerous tightly coupled applications running at 32K and 64K cores on the XT5.

NICS is specializing on true capability applications, plus high performance file and archival systems.

Typical job size on IB cluster at TACC is ~300 cores

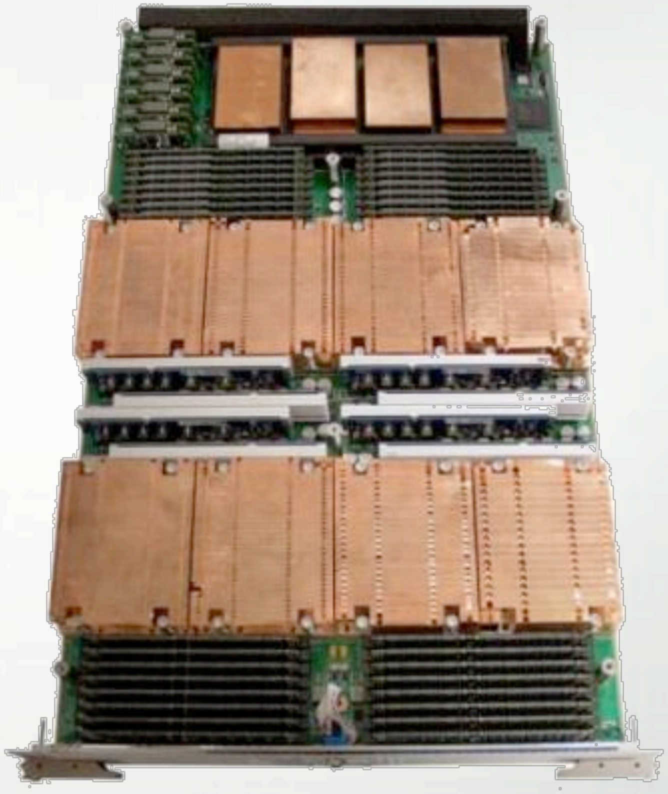
XT6 Processor Choices (vs. XT5)

Processor	Frequency	Peak (Gflops)	Bandwidth (GB/sec)	Balance (bytes/flop)
Istanbul (XT5)	2.6	62.4	12.8	0.21
MC-8	2.0	64	42.6	0.67
	2.3	73.6	42.6	0.58
	2.4	76.8	42.6	0.55
MC-12	1.9	91.2	42.6	0.47
	2.1	100.8	42.6	0.42
	2.2	105.6	42.6	0.40



Cray XT6 Compute Blade

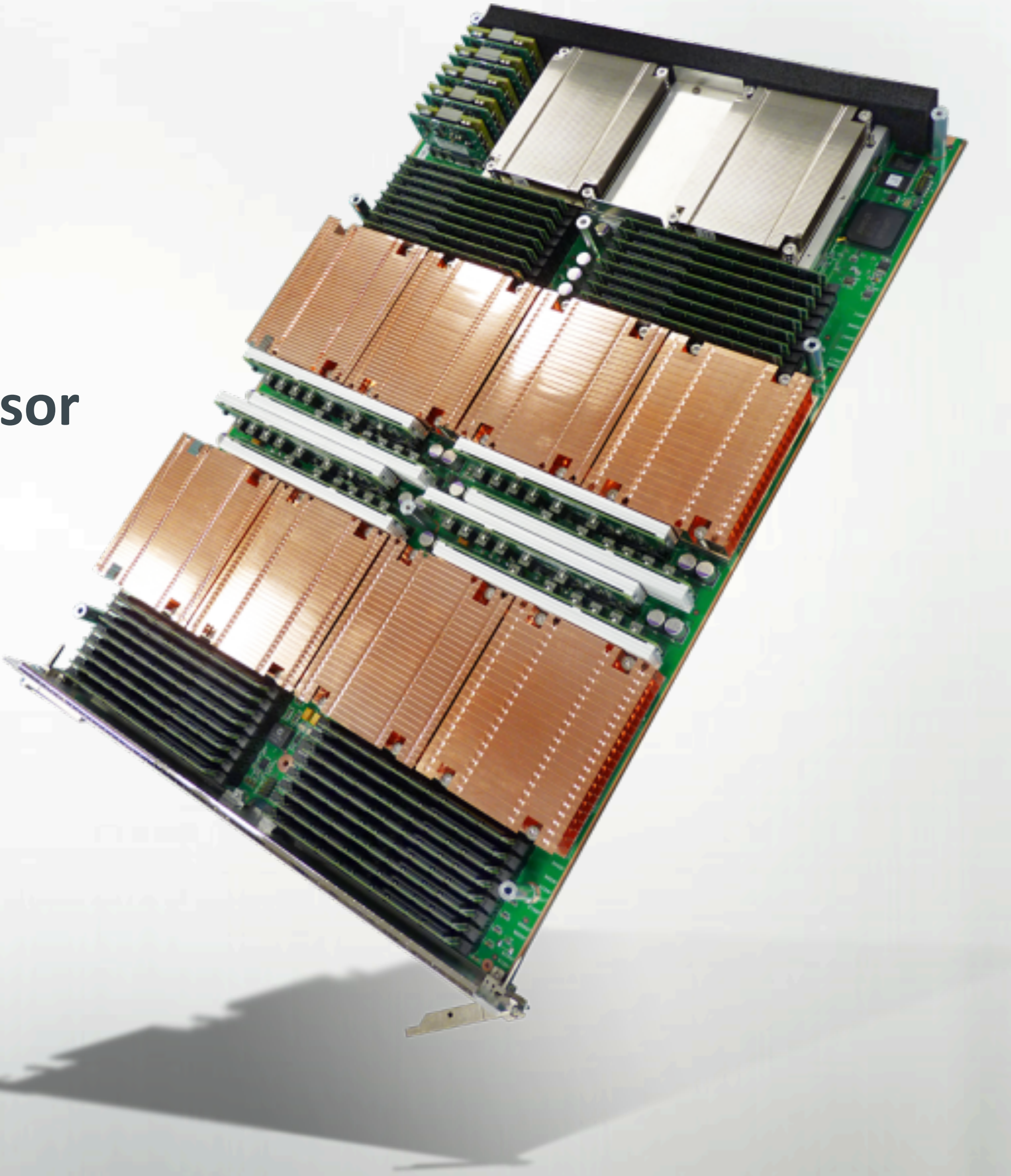
- New compute blade with 8 AMD Magny Cours processors
- Plug-compatible with XT5 cabinets and backplanes
- Initially will ship with SeaStar interconnect as the Cray XT6
- Upgradeable to Gemini Interconnect or Cray XE6
- Upgradeable to AMD's "Interlagos" series
- XT6 systems will continue to ship with the current SIO blade
- First customer ship, March 31st



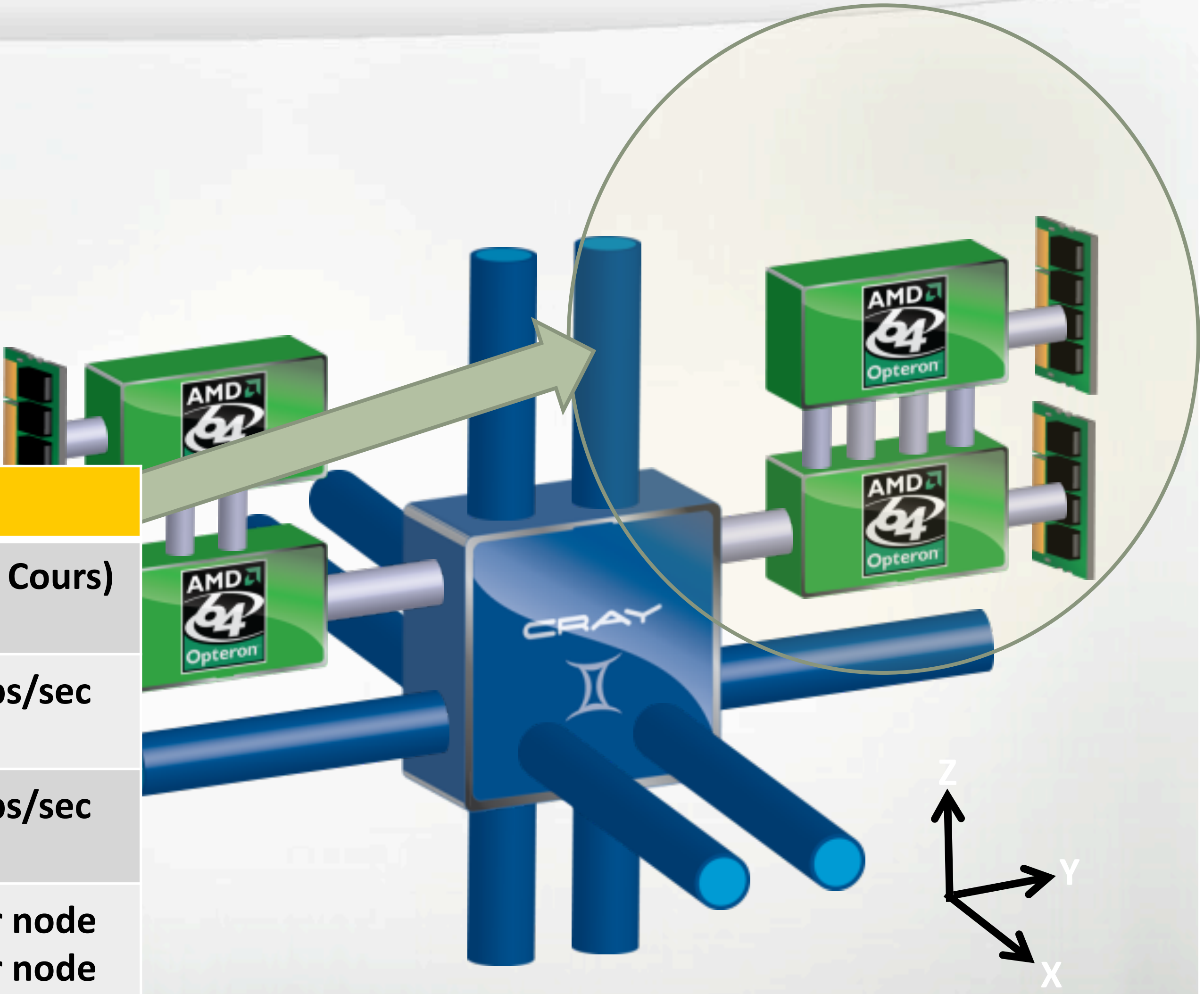
Cray XE6

Cray XE6 Compute Blade

- 8 Magny Cours Sockets
- 96 Compute Cores
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor

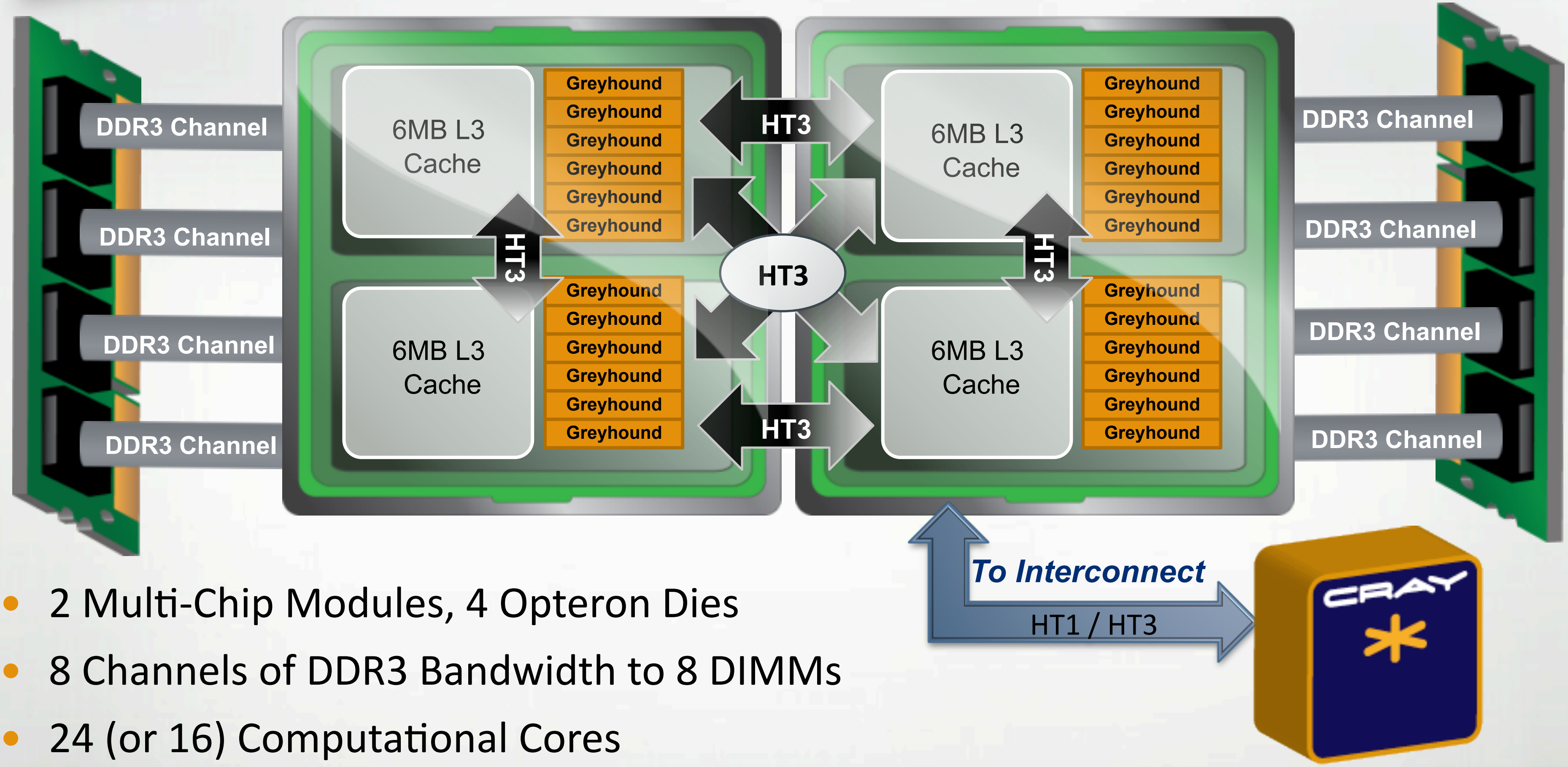


Cray XE6 Compute Node



Node Characteristics	
Number of Cores	24 (Magny Cours)
Peak Performance MC-12 (2.2)	211 Gflops/sec
Peak Performance MC-8 (2.4)	153 Gflops/sec
Memory Size	32 GB per node 64 GB per node
Memory Bandwidth (Peak)	83.5 GB/sec

XE6 Node Details: 24-core Magny Cours



- 2 Multi-Chip Modules, 4 Opteron Dies
- 8 Channels of DDR3 Bandwidth to 8 DIMMs
- 24 (or 16) Computational Cores
  - 64 KB L1 and 512 KB L2 caches for each core
  - 6 MB of shared L3 cache on each die
- Dies are fully connected with HT3
- Snoop Filter Feature Allows 4 Die SMP to scale well

Gemini Interconnect

Cray Network Evolution

SeaStar

- Built for scalability to 250K+ cores
- Very effective routing and low contention switch

Gemini

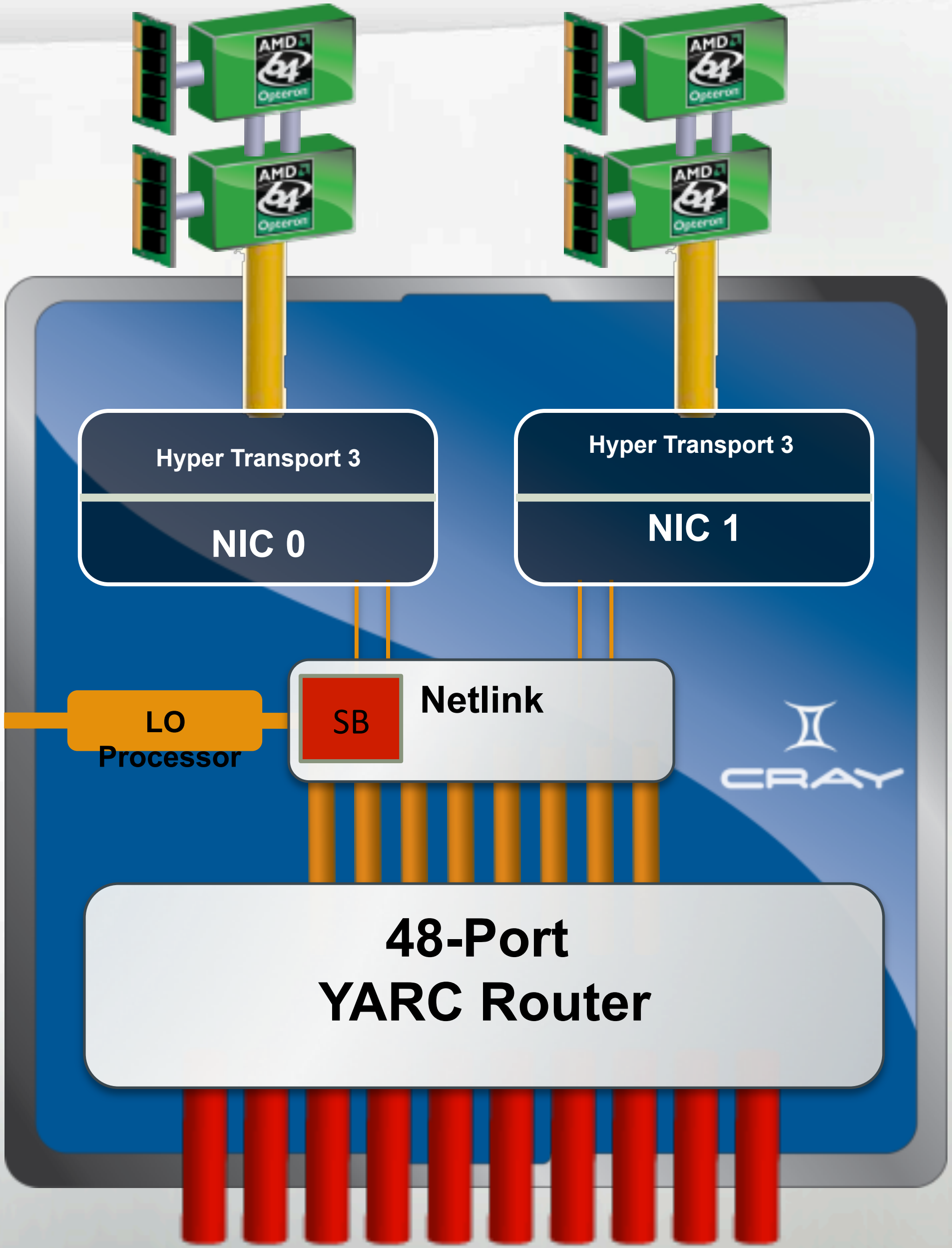
- 100x improvement in message throughput
- 3x improvement in latency
- PGAS Support, Global Address Space
- Scalability to 1M+ cores

Aries

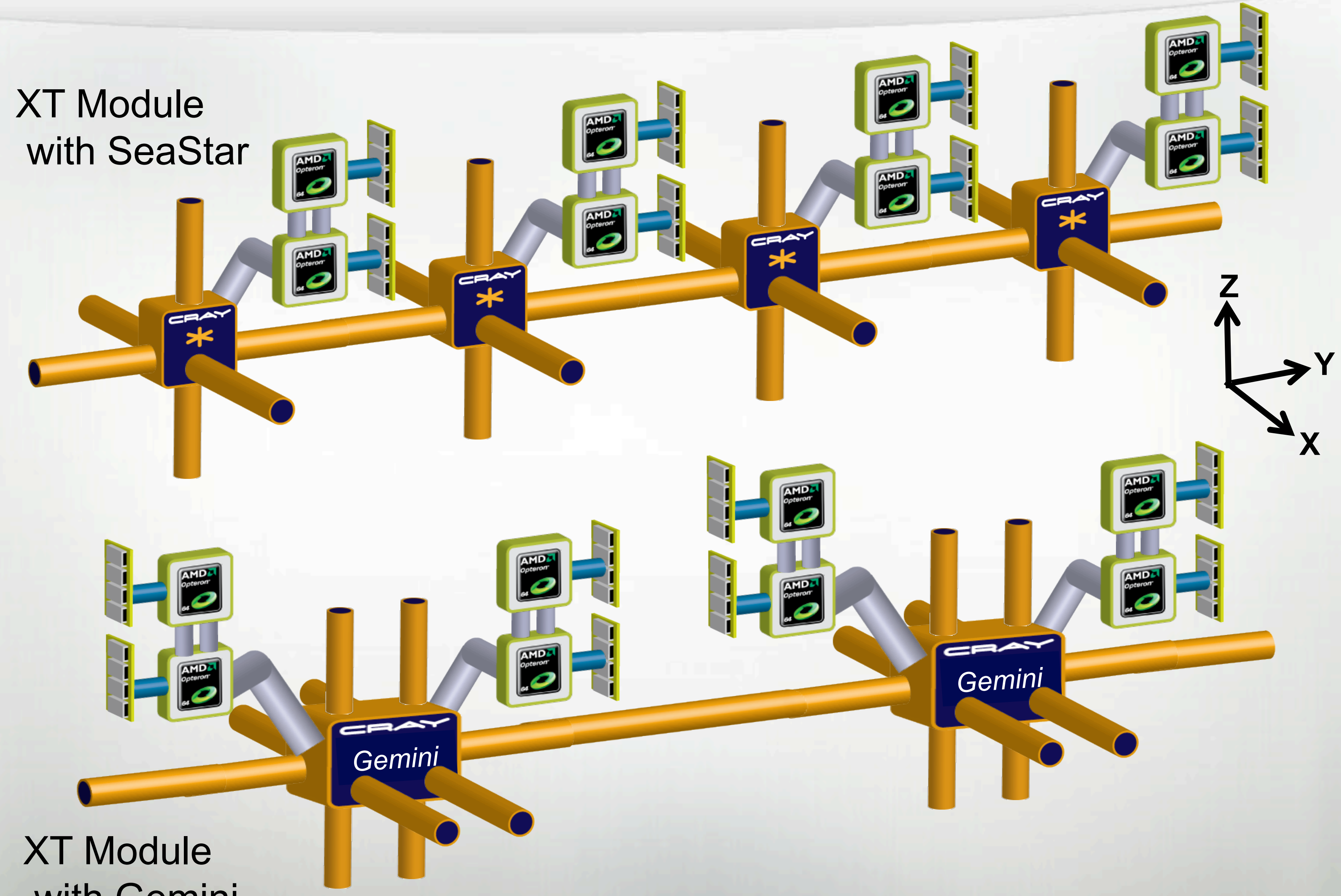
- Ask me about it

Cray Gemini

- 3D Torus network
- Supports 2 Nodes per ASIC
- 168 GB/sec routing capacity
- Scales to over 100,000 network endpoints
  - Link Level Reliability and Adaptive Routing
  - Advanced Resiliency Features
- Provides global address space
- Advanced NIC designed to efficiently support
  - MPI
    - Millions of messages/second
  - One-sided MPI
  - UPC, FORTRAN 2008 with coarrays, shmem
  - Global Atomics



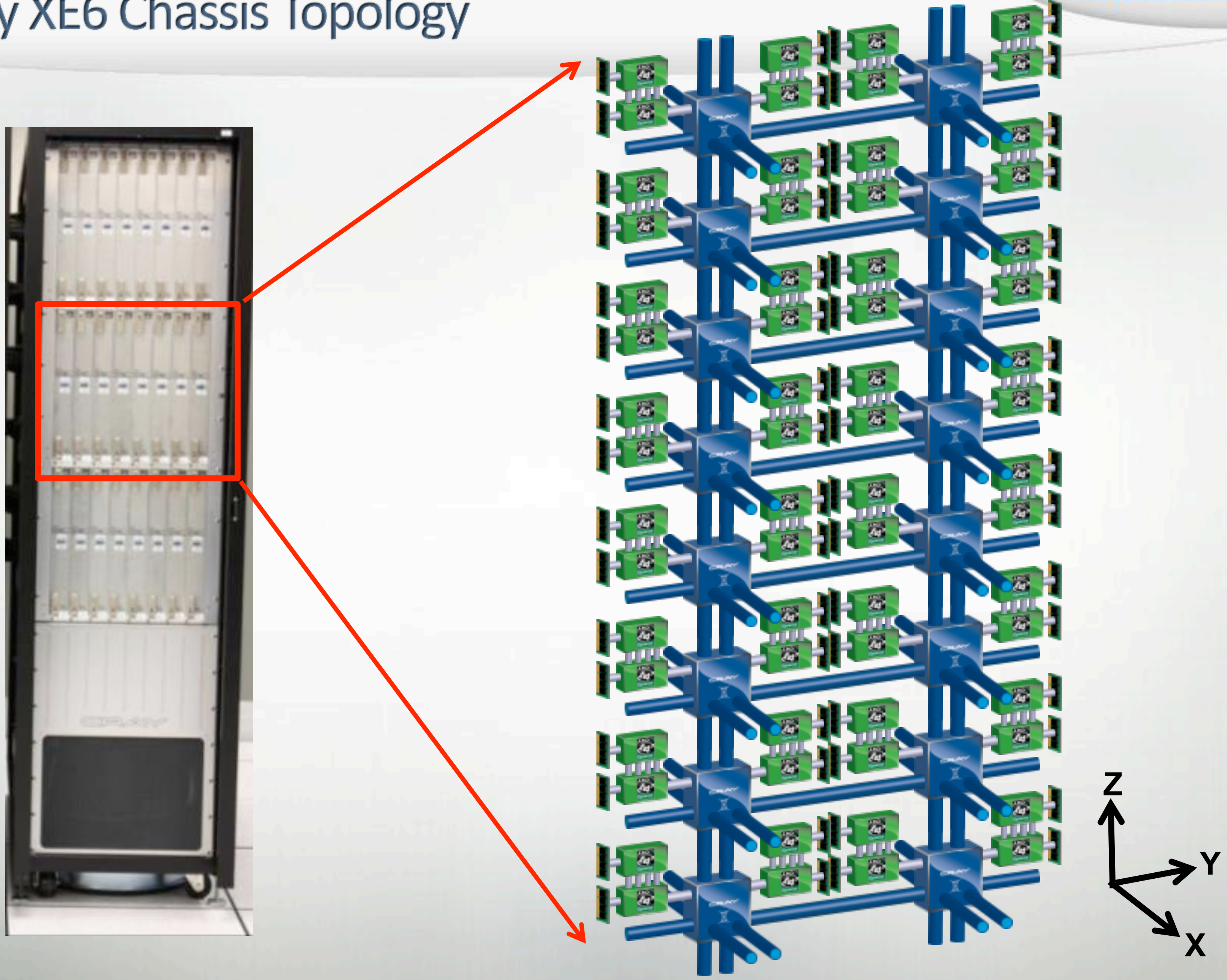
Gemini vs SeaStar – Topology



XT Module
with SeaStar

XT Module
with Gemini

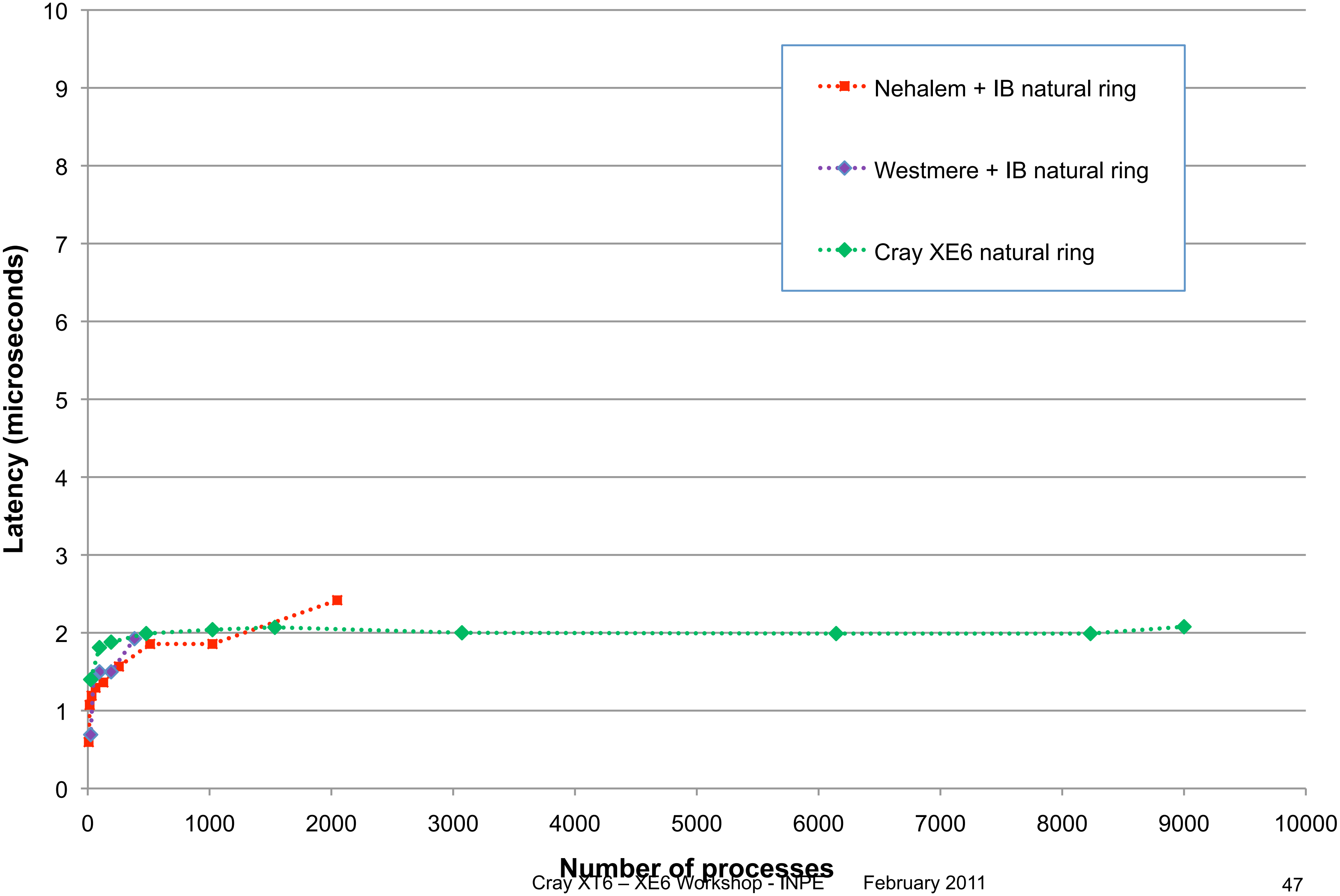
Cray XE6 Chassis Topology



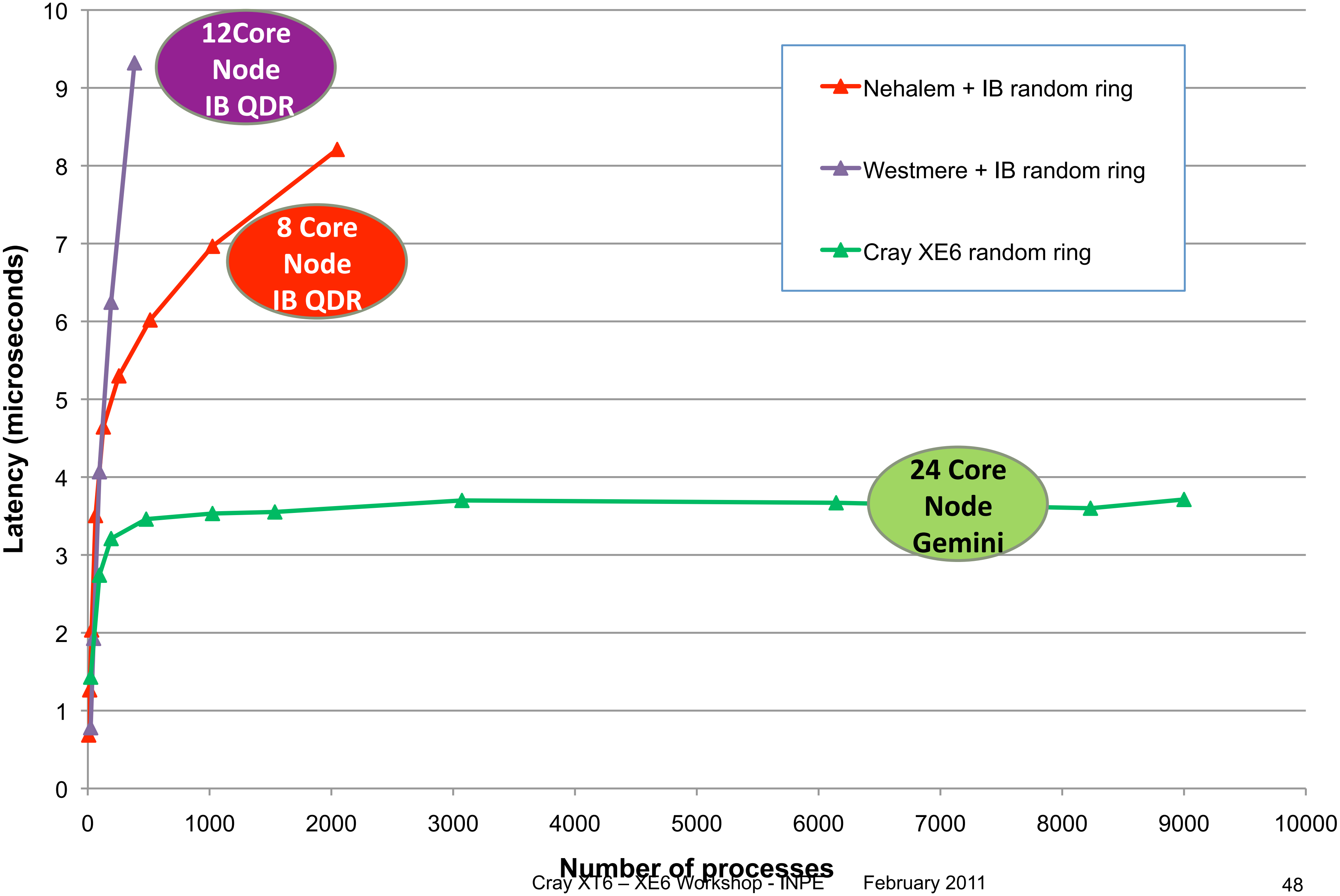
Gemini Advanced Features

- Globally addressable memory provides efficient support for UPC, Co-array FORTRAN, Shmem and Global Arrays
  - Cray Programming Environment will target this capability directly
- Pipelined global loads and stores
  - Allows for fast irregular communication patterns
- Atomic memory operations
  - Provides fast synchronization needed for one-sided communication models

Gemini – QDR Comparison
HPCC Natural Ring Latency Benchmark

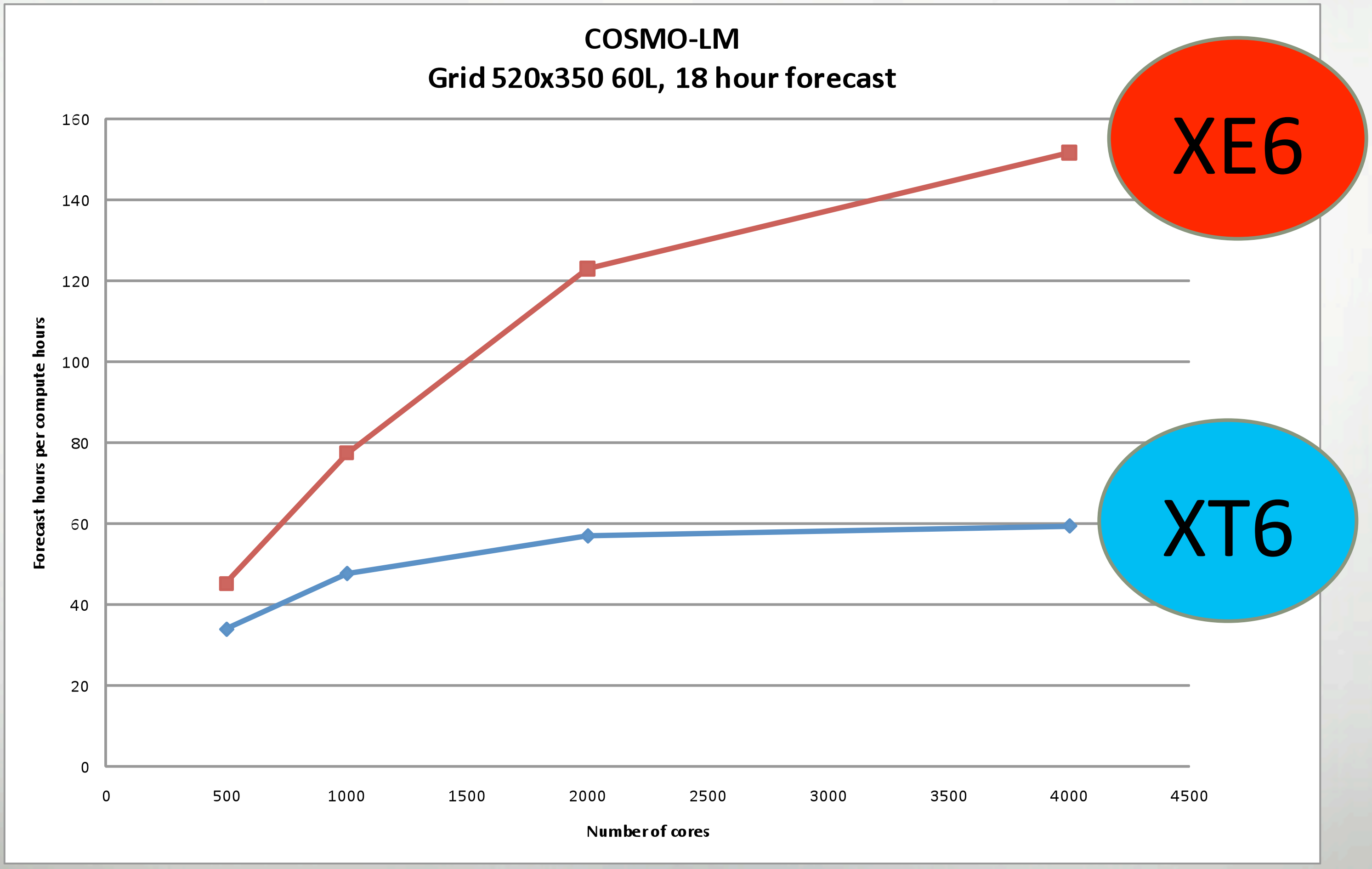


Gemini – QDR Comparison
HPCC Random Ring Latency Benchmark



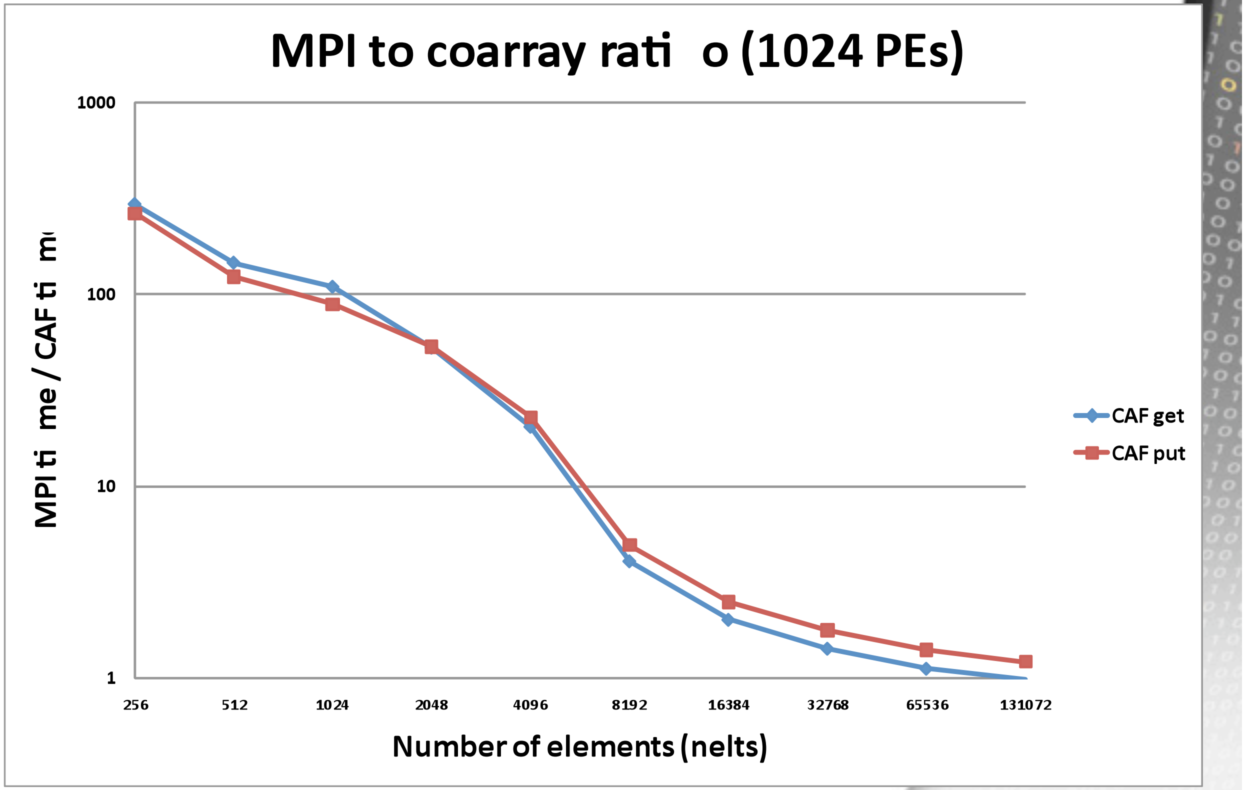
Scalability and simulation rate

- Forecast Hours per compute Hours
- Typical performance improvement



Remote gather: coarray vs MPI

- Coarray implementations are much simpler
- Coarray syntax allows the expression of remote data in a natural way – no need of complex protocols
- Coarray implementation is orders of magnitude faster for small numbers of indices



JAGUAR
CRAY XT5
APPLIED MECHANICS
Railway Technical Research Institute
H L R I S
CRAY XT5m
KRAKEN

Thank You!